A STUDY OF THE DATA MINING OF MEETING MINUTES OF CONSTRUCTION PROJECTS

by

Jaques van Niekerk

Thesis presented in fulfilment of the requirements for the degree of Master of Engineering in Civil Engineering in the Faculty of Engineering at Stellenbosch University

Supervisor: Prof. Jan Wium

December 2020

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save the extent explicitly stated otherwise), that reproduction and publication thereof by Stellenbosch University will not infringe any third-party rights and that I have not previously, in its entirety or in part, submitted it for obtaining any qualification.

December 2020

Jaques van Niekerk

Summary

This research is motivated by the increased use of big data and the need to decrease the cost/time overruns experienced in the construction industry. During the construction period of a project, numerous factors contribute to the outcome of the project. Simply knowing some of these factors may not contribute to the successful completion of the project. Being able to use the known and the unknown factors to create a model that can predict the outcome of a project will enable the project management team to make informed decisions.

This research aims to determine if the information currently being recorded in site progress meeting minutes, is sufficient to use in data mining applications for the prediction of the outcomes of a project, and to establish if new knowledge can be obtained from this process. Data mining to aid project management in the construction industry has seen limited application, especially in South Africa. Data mining is part of the Knowledge Discovery in Data (KDD) process, which is used to learn new information from data.

The research starts with a literature review to identify a list of factors that influence the outcome of projects – positively and negatively. From the identified project outcome factors, the two that are highlighted most often are leadership and planning. These two overarching categories were used to determine if and how influencing attributes are recorded in the site meeting minutes.

The current uses of data mining in the construction industry were investigated to determine how data mining and KDD have been implemented in the industry. Although KDD has been applied in the construction industry, no information was found about its application in the South African construction industry. Some of the reasons why it has not yet been implemented could be related to copyright, privacy and data security, and lack of incentives to implement data mining.

An investigation of several projects' meeting minutes was undertaken where the meeting minutes were data mined to determine if they can be used to predict the outcome of future projects. The two overarching categories above where used to identify the information that is present in the meeting minutes. These attributes were then used as the data mining features. Two data mining applications were used to compare the applications and to validate the results. The most accurate data mining models were created using the Random Forest data mining algorithm. The prediction models are able to predict the outcome of future projects with a high degree of certainty.

Opsomming

Hierdie navorsing is gemotiveer deur die toename in die toepassing van data in verskeie industrieë sowel as deur behoefte na suksesvolle projekte. Gedurende die konstruksie tydperk is daar talle faktore wat 'n rol speel in die uitkoms van 'n projek. Dit is belangrik om te weet wat hierdie faktore is en hoe hulle gebruik kan word saam met die onbekende faktore om die uitkoms van 'n projek te kan voorspel. Genoegsame data kan die projekbestuurspan in staat stel om ingeligte besluite te kan neem.

Die doel van hierdie navorsing is om te bepaal of die inligting wat tans in die projek se vorderingsvergaderings se notules aangeteken word, gebruik kan word om die uitkoms van projekte te voorspel sowel as om vas te stel of nuwe kennis gedurende hierdie proses verkry kan word.

Tot dusver was ontginning van data in die konstruksie bedryf van beperkte omvang in Suid-Afrika. Data ontginning is deel van die Kennisontdekking in Data (KDD) proses, wat gebruik word om nuwe inligting uit data te leer.

'n Literatuur oorsig is gedoen om 'n lys faktore te identifiseer wat die uitkomste van projekte beïnvloed – beide positief en negatief. Van die geïdentifiseerde projekuitkomsfaktore is die twee wat die meeste uitgelig word, leierskap en beplanning. Hierdie twee oorhoofse kategorië is toe gebruik om te bepaal hoe die atribute asook watter atribute, in die projek se vorderingsvergaderings se notules die uitkomste van die projek aanspreek.

Die huidige gebruik van data ontginning in die konstruksie bedryf is ondersoek om te bepaal hoe dit in die bedryf geïmplementeer word. Alhoewel KDD in die konstruksie bedryf toegepas is, kon geen inligting gevind word oor die toepassing daarvan in die Suid-Afrikaanse konstruksie bedryf nie.

'n Ondersoek na verskeie projekte se vorderingsvergaderings se notules is gedoen. Die notules is gebruik in die KDD proses, insluitend die ontginning van data, om te bepaal of die notules gebruik kan word om die uitkoms van toekomstige projekte te voorspel. Twee verskillende data-ontginningstoepassings is gebruik om die resultate te vergelyk en te valideer. Die mees akkurate data-ontginningsmodelle is geskep met behulp van die ewekansige woud algoritme. Die voorspellings modelle is in staat om die uitkoms van toekomstige projekte met 'n hoë mate van akkuraatheid te voorspel.

Acknowledgements

I wish to express my appreciation to the following persons. Without their support and assistance, this research would not have been possible.

- Prof. Wium, my study leader, who guided and assisted me in formulating and realising this research.
- Matthew Horwill who provided the data for this research.
- My family and loved ones for their constant support, encouragement and enthusiasm.
- RapidMiner for providing the software used for this research.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1. Introduction and background

The Project Management Book of Knowledge (PMBOK) says that a highly complex project environment is created when the speed at which information is distributed and decisions are made are coupled with project failure factors that might be present in certain projects (Project Management Institute, 2016). Chan and Kumaraswamy (1997) write that the majority of cost overruns typically occur during the construction phase in which many unanticipated factors are realised. The success and failure factors of projects have been widely published and discussed all around the world. In this research, the factors contributing to the outcome of a project are determined first. Following the identification of the factors, it is essential to ascertain whether they are being monitored during the execution of a project and how they are being recorded. By using Knowledge Discovery in Data (KDD), these factors can be data mined from existing and historic records to confirm their relevance.

The idea with performing data mining on stored information is to gain knowledge and insight, and to learn new interdependencies that might not have been known before. This new knowledge has the potential to assist a company in being more competitive in the market.

The project data can come from various sources and forms part of the tender documents. The original tender documents indicate the intent at the start of the project, as well as containing the specifications and other essential information. As the project progresses more information is generated through Requests for Information (RFI's), Variation Orders (VO's), meeting minutes, emails, schedules and other documents generated on site. Together these items contribute to the knowledge available on a specific project at a specific time. Applying KDD to this information could potentially influence the outcome of future projects. All construction projects are unique. Even if the same team constructs another project of similar type, it could potentially have a different outcome.

If the project management team has the ability to manage the project better by leveraging past project information, it can represent a competitive advantage which would ultimately lead to reduced project costs and more successful projects being completed. This research aims to investigate whether the meeting minutes of a project can be used to predict the outcome of future projects so the project team can act proactively to change the outcome of the project.

1.2. Research Aims and Objectives

This research aims to determine if the information, currently being recorded in site progress meeting minutes, is sufficient to use in data mining applications for the prediction of the outcomes of a project, and to establish if new knowledge can be obtained from this process. Data mining of site progress meeting minutes is investigated as part of a KDD process and is aimed at enabling project managers to manage their projects better. Recommendations are made on how project information should be recorded to support data mining processes in future. The research aim is achieved through realising the following objectives:

Objective 1: Identify factors that contribute to project success or project failure. Extensive research has been done on project success and failure factors. Through a detailed literature review, information will be collected from several case studies to determine the individual factors that contribute more to the success or failure of a project.

Objective 2: To determine the current extent of data mining application / use in the construction industry (as it has been applied in several other industries).

Objective 3: To investigate whether the factors in question are being recorded in the meeting minutes of a project and how the attributes can be turned into features by using the information obtained from Objective 1.

Objective 4: To apply the KDD method to existing site meeting minutes to identify new and relevant information. The objective is to determine whether the outcome of a project can be predicted with the features identified in Objective 3.

1.3. Scope and Limitations

Construction project management is a wide field with significant complexity. The same can be said of data mining. This study assumes familiarity with the civil construction industry in South Africa and the application of simplistic data mining concepts.

The project progress meeting minutes are a summary of the most important information on the project at a given time and will be used as the data source in this study. The meeting minutes vary from project to project, with some participants recording detail on site and others merely recording the overall progress on site. Due to the financial information of a project often being confidential, this information will be excluded from this study. Where financial information was recorded in the site minutes, it is excluded and is not transferred to the data warehouse.

The scope of the study is limited to the civil construction industry.

1.4. Methodology

This section briefly describes the methodology adopted to achieve the aim of the study within the defined scope and limitations. The methodology is described for each of the research objectives below.

Objective 1: The factors that contribute to success or failure of a project were determined through a literature review. Several case studies were obtained from accredited journals and publications and were used to condense the information identified by researchers into a concise list of factors to be used for the purpose of this research study.

Objective 2: The current status of the use/application of data mining in the construction industry was investigated through a literature study.

Objective 3: By using site meeting minutes from identified projects and by recording the data in a data warehouse, it was investigated whether factors identified under objective 1 are present in the meeting minutes that are used for this study. The list of factors that were identified was reduced to two overarching categories. The factors can be recorded objectively and/or subjectively by the parties in the way that the meeting minutes are kept. A recommendation is made on how the project attributes can be recorded to assist in the KDD process in the future.

An investigation of several meeting minutes was done. The site meeting minutes were sourced from a large construction company. Meeting minutes were taken from several projects, all of which are civil construction projects executed in South Africa over the last 10 years. The minutes form part of the historic contract documents that were archived by the construction company following the completion of the projects. Only site meeting minutes that have been archived electronically were used. Access to the minutes was granted by the managing director of the company with the provision that no customer or contractor names, nor any financial data may be used in the study. As mentioned above, financial information was ignored and was not stored in the data warehouse for the purposes of this research study.

Objective 4: During the investigation of several project meeting minutes, the KDD process was applied to the data gathered during objective 3. Machine learning algorithms were applied to determine if new and meaningful information can be obtained from previously recorded data.

Different data mining and machine learning techniques were evaluated to identify the most suitable techniques. All the data mining algorithms that are available in the data mining

applications were used to determine the best model(s) that can be created. The best model(s) are chosen based on the accuracy of the model after testing and validating the models. The process of data mining of the meeting minutes is described in Section 6.4. The methodology used for the investigation of the meeting minutes is discussed in Chapter 4.

1.5. Definition of key terms

Features: Features are used as headings in the data mining applications to recognise the different parameters that will be used to create the models. Features are also known as attributes, which is the terminology used by RapidMiner. For this study some project attributes that contribute towards the outcome of a project can be used as features for data mining as they are already in the correct format while other attributes need some transformations to be able to use them as features in data mining.

Data Mining Model: A data mining model for this research is defined as a logical and functional collection of objects.

Objects: The objects used in the data mining model are individual components that perform a role. For this study the objects that called widgets or operators depending on the data mining software being used, and are described in Section 6.4

Project Meeting Minutes: The meeting minutes being used for this study are the site meeting minutes. They are taken during progress meetings held between the client, contractor, consultant and other relevant parties. These meetings typically take place at least once a month or more often, as deemed necessary for a specific project. The meeting minutes normally contain all the information concerning the project, including start dates, projected end dates, progress, payment, difficulties, changes, safety, quality, equipment and weather.

Regression: Prediction values are continuous and cannot be placed into discrete values, for example a continuous range of numbers 1 to 100 (See Figure 1-1a).

Classification: Prediction values can be categorised into categories of discrete value, for example, yes/no/maybe (See Figure 1-1b).

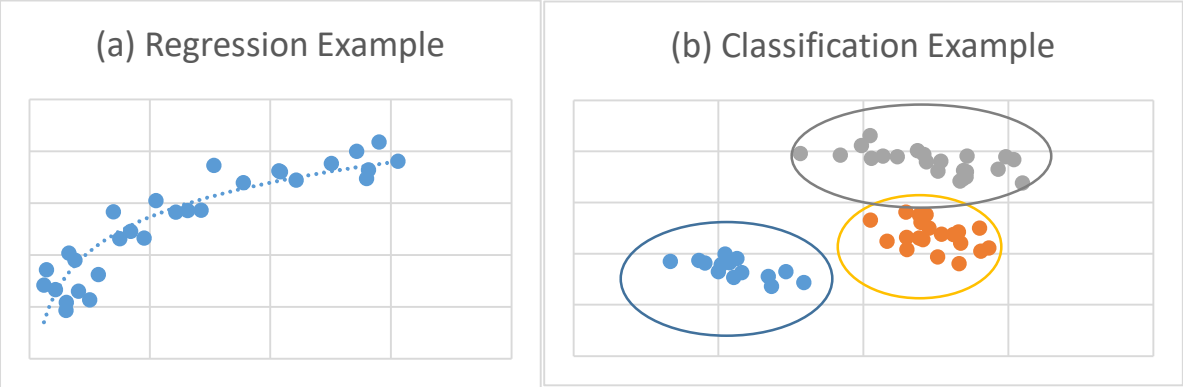


Figure 1-1: Explanation of Regression and Classification Problems

1.6. Research Outline

This section offers a brief layout of the research study with a short description of each chapter.

Chapter 1: Introduction

The introductory chapter comprises the introduction and background, research aim and objectives, scope and limitations, and methodology of this research study.

Chapter 2: Literature Review – Project Outcome Factors

The literature review is divided into two parts, each addressed in a separate chapter.

Chapter 2 investigates 10 case studies identifying the factors that contribute to the outcome of a project. Five of the case studies investigate factors contributing to the success of a project while the other five investigate factors contributing to the failure of a project.

Chapter 3: Literature Review: Knowledge Discovery in Data

Chapter 3 consists of four case studies that demonstrate the current status of data mining and/or KDD in the construction industry.

Chapter 4: Research Methodology and Data Collection

Chapter 4 discusses the research methodology adopted for the investigation of several project meeting minutes in Chapter 6. It provides a breakdown of how the data was collected and processed, and of how the investigation of several project meeting minutes was conducted.

Chapter 5: Analysis of the Most Relevant Success Categories

Chapter 5 takes the list of identified factors that contribute to the outcome of a project and narrows it down to two overarching categories. The two categories are discussed and the features that can be used in data mining are identified from them.

Chapter 6: Investigation of Several Project Meeting Minutes

Chapter 6 is an investigation into the data mined from the project meeting minutes of 27 civil construction projects completed in South Africa using the factors and features identified in the previous chapters. The investigation provides a practical example of how the records from past projects can be utilised to aid a project management team in identifying trends and in addressing them timeously.

Chapter 7: Proposal for Meeting Minutes

Chapter 7 provides a proposal on how to record the meeting minutes of a project to simplify the data warehouse creation.

Chapter 8: Conclusion

In Chapter 8, the findings of the research study are summarised with conclusions and recommendations.

2. Literature Review – Project Outcome Factors

2.1. Introduction

Different studies have looked into project success and failure factors in the past. According to Tshiki (2015), there seems to be an absence of consensus between the different researchers on the criteria and factors for judging the success of a project. Therefore, this research starts with the determination of the factors contributing to the success and failure of projects. A total of 10 case studies – five for successful projects and five for projects which failed – were reviewed to identify the project outcome factors. Because this is not an exhaustive list of literature, similarities were identified from various other literature sources that are mentioned within the various research studies. The goal of the literature review is to provide a list of factors that have an influence on the outcome of a project. It is also used to determine whether these factors are currently being recorded by reviewing the site meeting minutes of 27 completed projects.

2.2. Defining Project Success and Project Success Factors

There are numerous studies that have been done on project success factors. Chua et al. (1999) acknowledge the fact that project success factors vary for different project objectives. Ribeiro et al. (2013) indicate that in today's complex construction environment, projects require effective planning and communication with the help of advanced tools. These tools should enhance the quality of site management. Zulch (2016) states that ineffective communication on a project can lead to project objectives not being met.

Ribeiro et al. (2013) consider whether traditional factors are still relevant or whether other factors have become relevant for determining the success of a construction project. These traditional success factors are mostly related to the elements of the Atkinson Triangle (cost, time and quality). Based on a survey sent to 750 project managers, their research shows that project managers still consider the traditional factors of completing the project within budget and on time as the most important success factors. These two factors are followed by quality requirements and customer expectations. The feedback provided by the project managers is their subjective feelings on when and how a project can be considered successful (Ribeiro, et al., 2013).

These sentiments are shared by Ananthanarayanan and Devi (2017) in their article where they state that construction cost is the most important factor used in determining project success. Tshiki (2015) describes project success as an assessment of the overall objectives of the project and project management success as the assessment of the traditional performance factors (cost, time and quality).

David Pratt (2015) indicates that a good project manager should be able to provide an objective report on the status of the project. A firm grasp on the status of a project is only possible with adequate proof. Consequently, it is of vital importance to have several factors or indicators that one can apply to measure the health of a project.

There are several ways to determine the factors that influence the outcome of projects. Two of these methods are through literature studies and expert surveys. Most of the authors cited above have done extensive research on the factors that influence the outcome of projects. They use literature to develop and enhance the impact factors which were confirmed through surveys. The outcome of 10 of these research studies are used as case studies to establish a short-list of factors that influence the outcome of construction projects. These 10 research studies are divided into two equal groups: factors that contribute to the success of a project; and factors contributing to the failure of a project.

2.2.1. Research Study 01: Development of a Comprehensive Model for Construction Project Success Evaluation by Contractors

Heravi and Ilbeige (2012) have developed a comprehensive quantitative model for construction success evaluation from the contractor's point of view. They split the determination of the success of a project into two areas: project success based on (1) product success and (2) project management success. They argue that, even though a project can have substantial time and cost overruns, it can still be considered as a successful project. Thus, there is a need to split project management from the final product's success factors.

In their paper, Nguyen et al. (2004) point out that it is important to differentiate between project success and project management success. They continue to say that project success is determined by the overall objectives of the project, and project management success is determined by the Atkinson Triangle elements. Heravi and Ilbeigi (2012) mirror this

comment by stating that, by determining project management success, one is essentially disregarding product success. An example of this is the fact that one can manage a project successfully but, should the product delivered not meet the client's expectations or requirements, the project is ultimately unsuccessful. For this reason, the authors split their model into two distinctive sections.

Heravi and Ilbeige (2012) have done extensive literature reviews to determine the appropriate project success definition. From this definition they develop a model that consists of two parts: product success and project management success. The steps the authors followed in the development of the model are:

- Identification of the critical performance indices;
- Quantification of the performance indices;
- Normalisation of the indices;
- Integration of the various performance indices to develop product success and project management success functions; and
- Applying the indices to a model to test and analyse the results.

From extensive literature reviews, Heravi and Ilbeige have developed the following performance indices for construction companies' product success:

- Profitability performance index;
- Product quality performance index;
- Client satisfaction index;
- Contractor's professional profit satisfaction index; and
- Investment performance index.

From there, they proceeded to develop the following six indices for the project management success of a construction company:

- Cost performance index;
- Billing performance index;
- Schedule performance index;
- Safety performance index;
- Process quality performance index; and

- Environmental performance index.

After the identification of the various indices, Heravi and Ilbeige went through a process of normalising the indices to determine a product success function and a project management success function.

The authors applied the indices to major projects by an Iranian contractor who constructed power transmission lines. After highlighting the successes and failures of the project, the interpretation showed that, although the projects were successful and the contractor managed to make a good profit, the project management could be considered unsuccessful. Therefore, they would advise the contractor to proceed with caution, as this could certainly indicate a potential problem with the long-term success of a project.

2.2.1.1. *Summary*

The above research by Heravi and Ilbeigi (2012) can be used to determine the success of a project or of its project management when the relevant indices are available. Part of Objective 4 is to discover new knowledge through data mining. According to the Heravi and Ilbeigi, filling in the variables in an equation could indicate whether or not a project is successful. However, no new knowledge will be gained that can be used to benefit future projects or to make predictions about future projects. This case study is included as it provides valuable insight into the success factors of a construction project from a contractor's viewpoint. For the purposes of this research, the two groups of indices can be combined into a set of success factors for a project. The factors are (1) cost; (2) quality; (3) time management; (4) safety and (5) environmental performance.

2.2.2. Research Study 02: Factors in Project Success

The Association of Project Management (APM) have published a research report in which it develops a framework of project success by obtaining input from a literature review, interviews with professionals and academics, and the deliberations of APM and its partners. The factors identified are general factors relevant to project management and can relate to either the point of view of the client, the contractor or both.

APM asked 25 leading project management professionals and academics to define an initial framework of success factors. Then, they asked more than 850 project practitioners to validate and indicate the relevance of the framework success factors.

APM identified the following 12 project success factors that provide a framework for factors contributing to the success of a project:

- Effective governance;
- Goals and objectives;
- Commitment to project success;
- Capable sponsors;
- Secure funding;
- Project planning and review;
- Supportive organisations;
- End users and operators;
- Competent project teams;
- Aligned supply chain;
- Proven methods and tools; and
- Appropriate standards.

Each one of the main success factors has a small group of contributing factors. The 12 success factors (framework) were used as the basis for an online survey in which 862 project professionals participated. The respondents had to indicate how important each one of the factors is to project success. The main success factors had high averages of between 8.2 and 8.7 out of 10.

The main factors can be reduced to just six headline success factors. Figure 2-1 shows the relationship between the headline factors and the project success measures. The factors that are attributed to project success are project planning and review, clear goals and objectives, and effective governance. These factors occupy at least two of the top three places in each category.

Figure 2-1: Relationship between project success factors and headline success factors (Association for Project Management, 2014)

2.2.2.1. *Summary*

This case study is not specific to construction but to project management in general. The findings from the case study are, however, relevant to construction project management, and that is the reason this case study is included. Normally, in construction projects, much emphasis is placed on the three main factors (quality, cost and time) and little emphasis is placed on the three other, 'softer' factors (funder satisfaction, stakeholder satisfaction, and overall project success). This research has shown that it is important to look at a project holistically and to consider all the factors to determine if a project can be deemed successful.

For this study, it is important to note the following four factors:

- Project planning and review – Pre-project planning should be thorough with adequate monitoring and review throughout the project.
- Goals and objectives – The goal(s) of the project should be clearly defined to all stakeholders involved in the project.

- Effective governance – Clear reporting lines and regular meetings (good communication) should be a key cornerstone when setting up and during the execution of the project.
- Competent project teams – Ensure that the staff employed on the project are competent and have the correct mix of experience and skills required.

2.2.3. Research Study 03: Main Factors Influencing Project Success

Beleiu et al. (2015) present an overview of project success that they compiled through literature reviews and that they confirmed using quantitative research. The aim of their study is to identify the main factors contributing to the success of a project. Additionally, they try to find correlations between the factor considered having the greatest influence on project success and the other factors. According to the researchers, project success was initially described as reaching the objectives of the project in terms of time, cost and performance. However, they indicate that because the project management field has developed, the factors are no longer enough to define the success of a project.

The researchers developed a questionnaire that contained a list of 19 success factors where the respondents had to select the top five factors influencing the success of a project. The respondents then had to rank the statements on a Likert scale from 1 to 5 based on their experience. Based on the results from the questionnaire, the top five factors are:

- Clearly defined goal and direction (70.2%);
- Clearly defined roles and responsibilities (53.2%);
- Competent project team members (53.2%);
- Compliance with the planned budget, period and performance criteria (40.4%); and
- Communication and consultation with stakeholders (40.4%).

Coincidently, based on the number of times they were chosen by the respondents, the least important of the factors are owner involvement in the project and provision of timely data to the key players. The full list of success factors with their relevant percentages is illustrated in Table 2-1.

Table 2-1: Success factors by Beleiu et al. (2015)

Success factors	**Number of votes**	**Percentage of respondents choosing the factor**
Clearly defined goals and directions	33	70.2%
Competent project team members	25	53.2%
Clearly defined roles and responsibilities	25	53.2%
Compliance with the planned budget, timeframe and performance criteria	19	40.4%
Communication and consultation with stakeholders	19	40.4%
Accurate schedule and plan	17	36.2%
Synergy of the team	15	31.9%
Ability to handle unexpected problems	15	31.9%
Client acceptance of the results	11	23.4%
Timely and comprehensive control	10	21.3%
Adequate use of project management techniques	10	21.3%
Experience and expertise of the project manager	7	14.9%
Top management support	7	14.9%
Stakeholders satisfaction	6	12.8%
Adequate use of technical skills	5	10.6%
Adequate risk management	5	10.6%
Sponsor involvement within the project	3	6.4%
Provision of timely data to key players	2	4.3%
Owner involvement within the project	1	2.1%

2.2.3.1. *Summary*

The top five success factors highlighted by Beleiu et al. (2015) are all key components of the six headline factors as indicated by APM (2014) in section 2.2.2. The five factors mentioned in this research study can thus be treated as sub-factors, as they all underscore the importance of the 6 headline factors by the APM. These results are included in this study to support the importance of the factors that have been identified in Research Study 02.

2.2.4. Research Study 04: Critical Success Factors for Highway Construction Projects in Pakistan

The research done by Sohu, et al. (2018) involves a literature study where the authors determined 24 Critical Success Factors (CSF) and a questionnaire that was distributed to clients, contractors and consultants where the relevance of each of these factors was tested.

The objectives of the research were to (1) determine the CSFs of highway construction projects in Pakistan and (2) to rank CSFs of highway construction projects using their mean value.

The researchers did a literature review on the topic of project success factors to use in their survey. From their literature review, the researchers decided on the following CSFs for their questionnaire:

1. Experienced project management team;
2. Effective site management;
3. Effective communication and coordination;
4. Commitment of all parties to the project;
5. On time decision making;
6. Good relationship among project participants;
7. Experienced design team;
8. Availability of resources required for project;
9. Effective quality control system
10. Adequacy of raw materials;
11. Allocating appropriate funds;
12. Availability of equipment;
13. Capable supervisors;
14. Experienced sub-contractors;
15. Use of environment friendly equipment;
16. Support from senior management;

17. Implementation of safety management system;
18. Regular equipment maintenance;
19. Utilization of advanced technology;
20. Clear project goals;
21. Proper project planning;
22. Competitive procurement process;
23. Multidisciplinary project team; and
24. Correct estimation of project cost.

Sohu, et al. (2018) developed a questionnaire containing these 24 CSFs. They asked respondents to provide a score between 1 (Not Significant) and 5 (Extremely significant) for each of the 24 factors in terms of their significance as a CSF in a highway construction project in Pakistan. The questionnaire was distributed to 250 construction professionals and 130 completed questionnaires were collected. The respondents were from the client (38%), the contractor (34%) and the consultant's (28%) organisations. The top-seven ranking responses, which all received an average score of 4 out of 5 and above, were (1) experienced project management team(, 2) effective site management, (3) commitment of all parties to the project, (4) experienced design team, (5) proper project planning, (6) effective communication and coordination, and (7) allocating appropriate funds.

2.2.4.1. *Summary*

The findings from Sohu, et al. (2018) are limited to the construction of highways in Pakistan. The highest-ranking CSFs, that have been identified by the authors, can be used as factors in other construction projects in general. Pakistan and South Africa are both developing countries and, thus, the CSFs could be transferable to projects in South Africa. The above-mentioned research is, therefore, relevant to this study as it effectively provides a list of success factors that are relevant to South Africa.

2.2.5. Research Study 05: Critical Success Factors Influencing Project Success in the Construction Industry

The study by Garbharran and Govender (2012) is based on the four COMs model (comfort, competence, communication and commitment). Garbharran and Govender conducted a survey among 95 project managers and 61 active grade-four (CIDB rated) contractors in Durban, South Africa. The researchers set out to determine to what degree the project

managers and contractors agreed with the four COMs model's critical success factors in achieving project success.

The four categories of the four COMs model identify and group various success factors. Each one of these categories is explained below. The findings from the authors' research are also explained within each sub-heading.

The questionnaires were sent out to project managers and contractors in the commercial construction industry (excluding residential construction) from both the private and public sectors.

2.2.5.1. *Comfort*

This group emphasises that successful projects always involve primary and secondary stakeholders. It is therefore important to have a competent project manager as well as a resource management plan. This is to ensure that competition for resources during the project is managed correctly. Some of the other factors in this group include adequate funding for the project and comprehensive contract documentation.

The researchers find that at least 80% of the respondents felt that the factors that constituted comfort were "extremely important". Having comprehensive contract documents was ranked the highest from the respondents but interestingly, contractors ranked this item the lowest. Table 2-2 indicates the ranking of the various factors from the responses by the project managers and the contractors.

Table 2-2: Ranking of comfort factors by respondents

Factor	**Ranking by project managers (n=95)**	**Ranking by contractors (n=91)**
Comprehensive contract documentation	1	5
Involvement of stakeholders	2	1
Competent project manager	3	3
Adequate funding	4	5
Availability of resources	5	2

2.2.5.2. *Competence*

The following four factors form part of the competence group: Use of up-to-date technology, emphasis on experience, competent teams in place, and the aspects of how bids are awarded. These factors are critical for the success of a construction project. Using technology and experience to benefit the project is self-explanatory. Consideration should be given when awarding the bid to a specific project manager or team. A skill analysis Should be done to ensure that the team and/or project manager is suited to the specific type of project. The ranking of the competence factors by the two groups of respondents is shown in Table 2-3.

Table 2-3: Ranking of competence factors by respondents

Factor	**Ranking by project managers (n=95)**	**Ranking by contractors (n=91)**
Awarding bids to the right project manager/contractor	1	3
Use of up-to-date technology	2	1
Proper emphasis on past experience	4	2
Competent project team	4	4

2.2.5.3. *Commitment*

For any project to be successful, must be buy-in and commitment to the project from the project team, and there must be support from top-level management. Other important factors in this group include clear objectives and political support.

According to Garbharran and Govender (2012) over 80% of the respondents considered the factors contributing to the commitment group to be “extremely important”. The ranking of all the factors that form part of the commitment group by the two groups of respondents is shown in Table 2-4. A clear dissimilarity is evident in the fact that the project managers considered political support as the highest-ranking factor, whereas contractors considered top management support as the highest-ranking factor. Political support is regarded as important by contractors and most of the construction project public sector.

Table 2-4: Ranking of commitment factors by respondents

Factor	Ranking by project managers (n=95)	Ranking by contractors (n=91)
Political support	1	2
Clear objectives	1	4
Commitment to the project	3	3
Top management support	4	1

2.2.5.4. *Communication*

Successful projects need a shared project vision with frequent project meetings that involve the community where needed. The project documents should be updated regularly (in other words. drawings), and the handover procedures should be well communicated.

According to Garbharran and Govender (2012), at least 71.6% of the respondents indicated that communication is "extremely important". Both project managers and contractors indicated that handover procedures are the most important factor in this group. According to the authors, the importance of handover procedures lies in the fact that the industry is moving away from working for the client to working with the client (in other words, it is becoming a service industry). The ranking of all the factors by respondents in this group is shown in Table 2-5.

Table 2-5: Ranking of communication factors by respondents

Factor	Ranking by project managers (n=95)	Ranking by contractors (n=91)
Handover procedures	1	1
Frequent project meetings	2	4
Regular update of plans	3	5
Community involvement	3	2
Shared project vision	5	2

2.2.6. Summary

Garbharran and Govender (2012) find that the adequacy of the plans and specifications, and the competency of the project managers were two of the most important success factors when it comes to project success. This is substantiated by the findings from a research paper by Kog and Loh (2012). The 14 factors that have been divided into four groups based on the four COMs model are all indicated as "extremely important" by the respondents of the study. This is a South African case study performed with South African contractors and clients, making this an important addition to this research.

2.3. Discussion of Findings (Success Factors)

The case studies above are chosen specifically to create a list of factors that influence the outcome of successful construction projects. Although some of the case studies are not specifically related to construction, the factors presented are relevant to construction projects. The success factors are summarised in Table 2-6. From the case studies concerning successful projects, the most important factor (mentioned by four out of the five case studies is budget (cost). The remaining four are ranked from the next most important to least important: (1) time management; (2) communication and consultation with stakeholders; (3) competent team members and (4) goals and direction (three out of five case studies supported these findings.)

Table 2-6: Summary of success factors

	Budget (cost)	Specification & quality	Time management	Safety	Environmental factors	Stakeholder satisfaction	Funder satisfaction	Overall project success	Communication and consultation with stakeholders	Clearly defined roles and responsibilities	Competent team	Goals and direction
Heravi & Ilbeigi (2012)	X	X	X	X	X		X					
Association for Project Management (2014)	X	X	X			X	X	X				
Beleiu et al. (2015)	X		X			X			X	X	X	X
Sohu et al. (2018)	X								X		X	X
Garbharran & Govender (2012)									X	X	X	X
	4	2	3	1	1	2	2	1	3	2	3	3

2.4. Reasons Why Projects Fail

Calleam (2019) states that he could mention more than 101 reasons why projects fail. These factors are general factors applicable to any type of project. The factors Calleam points out can be broadly summarised into 12 categories:

- Goal and vision;
- Leadership and governance;
- Stakeholder engagement issues;
- Requirement issues;
- Estimation;
- Planning;
- Risk management;
- Architecture and design;
- Information management;
- Quality;
- Project tracking; and
- Decision making problems.

The next set of case studies evaluates these and any other factors that the authors identified.

2.4.1. Research Study 06: Cost Overruns and Failure in Project Management

Doloi (2013) looks at a range of factors that affect the cost of construction projects. His research involves the investigation of the perceived understanding and the importance of the key attributes of achieving the target-cost performance. He views these factors from three perspectives – that of the contractor, client and consultant (all of whom worked on Australian projects). He indicates that several generic research papers have been published regarding the search for the causes of cost performance issues. However, little research has been reported in the context of understanding the root causes.

A questionnaire was sent to 160 construction clients, consultants and contractors. A total of 94 responses were received (29 consultants, 24 clients and 41 contractors). Prior to the questionnaire being sent, out Doloi conducted feedback sessions with five senior industry professionals on the list of factors that he put together based on his literature review. The final questionnaire consisted of a list of 73 factors where the respondents were requested to provide their objective opinions on how vital the factors are that impact project cost performance. From these 73 factors, 48 of them had a sizeable correlation and were further used in the study. The 73 factors were grouped together in seven groups:

- Project related;
- Contract related;
- Project management team related;
- Quality related;
- Planning related;
- Market-related; and
- Contractor related.

The 48 factors were reduced to 36 and then further reduced to eight. The importance and relevance of these eight factors were confirmed by regression analysis. Doloi performed multivariate regression analysis to produce the five most significant factors that could influence the cost of a project. These factors were accurate project planning and monitoring, effective site management, contractor's efficiency, design efficiency and communication.

2.4.1.1. *Summary*

This case study is included because it focuses specifically on some of the factors that are highlighted by the case study in Section 2.3 as being key elements of project success. These factors are project planning and effective site management. Through Doloi's aim to investigate the factors that influence the cost of a construction project, he re-iterates the importance of planning and effective leadership on a construction project.

2.4.2. Research Study 07: A Comparative Study of Causes of Time Overruns in Hong Kong Construction Projects

The research by Chan and Kumaraswamy (1997) focuses on the views of industry participants regarding the causes of time overruns of local projects in Hong Kong. This was achieved through a survey and interviews. The study also tries to differentiate between the perceptions of the three main participants in a construction project (clients, consultants and contractors).

The authors conducted a survey where they used 83 previously identified delay factors that were grouped into eight categories. They asked respondents to list factors that cause project time overruns and, then, to list these factors in order of importance. Respondents were invited to add additional factors they felt contributed towards project delays.

The 83 factors were hypothesised by the authors based on a survey (and subsequent interviews) in a previous study. These factors were grouped into eight major categories by the authors:

1. Project-related;
2. Client related;
3. Design team related;
4. Contractor related;
5. Material;
6. Labour;
7. Plant and equipment; and
8. External factors.

The authors used the relative importance index method to determine the ranking of different causes. This was transformed from a five-point scale that was subsequently used in the survey document. It was important to compare the relative importance of factors across the three groups of respondents, because each group would, naturally, be looking at

the failure factors in a project from a different viewpoint. From this analysis, the authors determined that, at that time in both building and civil engineering projects in Hong Kong, the five most significant factors for project delay were (1) poor site management and supervision; (2) unforeseen ground conditions; (3) low speed of decision making involving all project teams; (4) client-initiated variations; and (5) necessary variations of works. It is important to note that all three parties noted poor site management and supervision as being a very significant factor contributing to project delays.

2.4.2.1. *Summary*

Although this research was done in 1995, it is still relevant today. The importance of scheduling and proper planning prior to execution, followed by monitoring during construction has been highlighted in this research. Site management and supervision play an important role in the outcome of a project. The authors emphasise the importance of projects finishing on time and stress that adequate planning from the beginning of the project, even before actual construction takes place, is important in limiting cost and delay overruns.

2.4.3. Research study 08: A Reflection on Why Large Public Projects Fail

This research by Holgeid and Thompson (2013) is related to large information technology (IT) projects that fail in the United Kingdom (UK). The authors studied contextual issues and ended up revealing two core dimensions from which they developed a new framework.

Holgeid and Thompson conducted a literature review finding that there are various reasons cited for project failure. From this literature review, the authors selected literature that, in their opinion, was relevant to large public service projects. The key research questions asked by the authors are:

1. To what extent do large IT projects fail?
2. Why do IT projects fail?
3. What can be done about it?

The selected literature was presented in four categories: (1) contextual factors; (2) diagnosing failure; (3) understanding failure; and (4) preventing failure.

2.4.3.1. *Contextual Factors Influencing Public IT Project Failure*

Holgeid and Thompson discussed several views from their literature study. Some of these are generic, such as the lack of skills and leadership from senior management, while others are specific to the IT environment, such as the use of open standards and architecture. All of the literature indicated project size to some extent. A list of reasons was compiled of why, in their opinion, public IT projects are risky.

2.4.3.2. *Diagnosing Failure*

This section primarily focuses on answering Holgeid and Thompson's (2013) first question: To what extent do large IT projects fail? From the literature study, the authors quote specific reports that are widely recognised in the industry – the Standish Group Chaos (1994) reports on IT failure which is published every second year. Some of the findings from the report are shown in Table 2-7. The Standish group defines project success as "the project is completed on time and on-budget, with all features and functions as initially specified". They have two categories for project failure. The first category is described as a challenged project and the second category is described as an impaired project. The definition of a challenged project is "a project that is completed and operational but over budget, over time estimate and offers fewer features and functions than originally specified". A project is considered impaired when "the project is cancelled at some point during the development cycle". Holgeid and Thompson indicate that the Standish reports are being debated by academics and practitioners as to the exact extent of the failures, but that it seems all agree that IT projects have a significant failure rate as indicated by the findings from the literature review.

Table 2-7: Findings from Standish Group research (Holgeid & Thompson, 2013)

	1994	**2009**
Projects succeeded:	16%	32%
Projects failed:	31%	24%
Projects challenged:	53%	44%

2.4.3.3. *Understanding Failure*

In this section Holgeid and Thompson set out to answer their second question: Why do IT projects fail? Holgeid and Thompson look at the impact of the size, volatility, over-ambitious and complex projects, lack of skills, the black swan effect, management, leadership and technical factors. They discuss each of these based on their literature review. The management and leadership factors are of importance to this study. Holgeid and Thompson find the following management causal factors to be relevant:

- Poor leadership in project delivery;
- Poor stakeholder communication;
- Poor competencies (and skill shortages);
- Poor stakeholder management;
- Poor estimation methods;
- Poor risk management; and
- Insufficient management support.

2.4.3.4. *Preventing Failure*

This section of the report answers Holgeid and Thompson's third question: How can failures be avoided? The authors discuss the following topics: (1) early detection of potential failure; (2) IT projects as change journeys; (3) risk management strategies; and (4) management principles associated with IT project success.

There are 12 dominant early warning signs according to the literature. These can be divided into people-related and process-related risks (Table 2-8). The authors discuss how IT projects can be considered a journey that constantly changes, where an organisation moves from the current state to a desired state. Holgeid and Thompson highlight some factors from their literature review about risk management strategies and note the importance of recognising the hidden risks that are difficult to avoid or mitigate in a planned fashion. For this, they recommend the "black swan" stress test. They complete the study with a discussion on their findings and a conclusion where possible future work and limitations on the study are discussed.

Table 2-8: Early warning signs for preventing project failure (Holgeid & Thompson, 2013)

Most influential people-related risks	Most influential process-related risks
• Lack of top management support. • Weak project manager. • No stakeholder involvement and/or participation. • Weak commitment of the project team. • Team members lack requisite knowledge and/or skills. • Subject-matter experts are overscheduled.	• Lack of documented requirements and/or success criteria. • No change-control process (change management). • Ineffective schedule planning and/or management. • Communication breakdown among stakeholders. • Resources are assigned to a higher priority project. • No business case for the project.

2.4.3.5. *Summary*

Although this case study is about the failure of large IT projects in the public domain, it is relevant to this is study as many of the failure factors are present in all project management fields.

After doing a narrow and focused literature review, Holgeid and Thompson (2013) find that leadership-related issues are the over-arching theme mapped as a key factor by 50% of the literature. They also mention two contextual factors that could influence the success or failure of a project, namely size and volatility. There are several other factors that they mention in their literature review that are relevant to this study. These factors are mentioned as part of the people- and process-related risks. Some of these factors include leadership-related issues, and communication and planning-related issues. Most of the above-mentioned factors are considered contributing factors with regards to the project outcome.

2.4.4. Research Study 09: Construction Project Failures Factors in Vietnam

Nguyen & Chileshe (2014) use a mixed-method approach in their study. The study is compared to the BRICS and CIVETS construction environments. The BRICS countries are Brazil, Russia, India, China and South Africa, while the CIVETS countries are China, Indonesia, Vietnam, Egypt, Turkey and South Africa. The aim of their study is to evaluate perceptions of professionals in the construction environment regarding the critical factors leading to the failure of construction projects in Vietnam.

Nguyen and Chileshe's (2014) literature study presents an overview of risk-management practices and a summary of project failure factors in Vietnam and selected BRICS and CIVET countries.

In their literature study that includes eight case studies, Nguyen and Chileshe find nine common critical factors causing failure in construction projects. These nine critical factors were mentioned in four or more of the case studies. These factors, in order of their relative frequency of occurrence, are (1) poor design and frequent design changes; (2) financial difficulties of the contractor; (3) obsolete or unsuitable construction methods; (4) incompetence of the construction team; (5) poor site management and supervision; (6) slow progress payments; (7) financial difficulties of the client; (8) corruption; and (9) bureaucratic administrative system. Eleven additional factors are mentioned, but most of these are mentioned by only two of the eight case studies.

Nguyen and Chileshe (2014) set out to perform a questionnaire survey where they compared the relative importance of each of the factors. They validated the questionnaires with interviews. Nine out of the ten interviews were with interviewees who were involved in the construction industry, while the tenth interviewee was an academic.

The results from the survey led to the conclusion that there are three main factors that influence the success of a construction project in the BRICS and CIVETS construction environments. The factors can be summarised as poor planning, lack of experience and frequent design changes. Other factors that contribute to the failure of construction projects to a lesser degree are lack of knowledge or ability to manage construction projects, lack of financial capacity of owners, and poor performance of contractors and sub-contractors.

The authors conclude with recommendations for practitioners, government and academia in Vietnam. These recommendations include strategies:

- For effective project planning processes;
- For the improvement of knowledge for executing complicated projects;
- Relating to design capacity; and
- To increase awareness, adoption and use or risk-management practices.

2.4.4.1. *Summary*

This case study is of importance to this research as it was carried out in a country experiencing similar construction tendencies to South Africa. The research by Nguyen and Chileshe (2014) also points to another factor that could be used in data mining site meeting minutes, namely frequent design changes.

2.4.5. Research Study 10: The Paradox of Time and Cost Overruns: A Review of Literature in Developing Countries

Sitwala and Wium (2013) argue in their research paper that, although many attempts have been made to solve the problem of time and cost overruns on projects, it appears that no solution has been found and, thus, the paradox. The paper reports on the preliminary findings of a research project that investigates the management of large projects in developing countries through a literature review into project over-runs.

The authors start off by indicating that a total of 25 studies were used for the purposes of their research paper, with 12 studies investigating successful project management, eight investigating time and cost-overruns in projects and five investigating delays in projects. Most of the studies were concerning projects in Nigeria, followed by South Africa, Libya, Egypt, Ghana and Zambia.

Sitwala and Wium (2013) notes that different authors provide different information on the extent of project over-runs. Some refer to the numbers of projects affected while others refer to the percentage of over-runs suffered on projects. However, most of the authors focused on the causes of over-runs rather than the extent; therefore, the statistics they provided vary significantly.

Sitwala and Wium (2013) highlight the probable causes of time and cost over-runs in projects in the Middle East, Africa and Asia. This study focuses on the causes they highlight for Africa. They indicate that the causes are similar in all three regions, with the

primary cause being the overdue payments to the contractor, followed by the financial difficulties of the contractor. Other causes that are mentioned include poor pre-planning resulting in frequent change orders from clients; use of procurement systems which results in poor contract management; bureaucracies when dealing with governmental departments; and the contractor being blamed for inefficient contract, materials and labour management systems.

In their conclusion Sitwala and Wium (2013) also mention the causes of project over-runs caused by the consultant organisations. The problems seem to result from incomplete drawings and documents and poor coordination of projects.

Sitwala and Wium (2013) surmise that the problem of over-runs are rooted in the management process of construction projects because they are divided into design and construction with the two parties on opposite ends. However, over-runs are also affected by the actions of other players outside of the project, including suppliers, personnel trainers and government, especially in large projects. Finding solutions to harmonise the players and to integrate them could be a way of beginning to solve the problems.

2.4.5.1. *Summary*

This case study is relevant to this research as it focuses on some of the important factors that can influence the outcome of projects in South Africa. The factors that were identified include planning, contractor delays and client-initiated variations.

2.5. Discussion of Findings

The factors presented in Research Studies 6 to 10 are not just relevant to the construction industry, but have been cited in other research about failure factors in projects. Moyle (2014) writes that, in his experience, projects rarely fail because somebody did something wrong halfway through the project. He finds that the failure of projects can mostly be attributed to a lack of clarity and objectivity at the definition stage. Marr (2016) lists seven reasons he feels tech projects fail. These factors overlap with the factors that Calleam (2019) lists. Marr (2016) states that poorly-defined outcomes, lack of leadership, lack of accountability, insufficient communication, no planning, lack of user testing and solving the wrong problem are all factors that have an impact on the outcome of a project. Not defining the problems that must be solved and how these will be solved can lead to the

project being defined as a failure because it does not meet the client's requirements. This, however, is not as relevant for construction projects, but poorly-defined outcomes and solving the wrong problem both fall under the same broad category. Similarly, a lack of accountability by executives falls into the same category of lack of leadership.

Table 2-9 provides a summary of the literature review on reasons why projects fail.

Table 2-9: Summary of failure factors by author

	Goal and vision	**Leadership and governance**	**Client-initiated variations**	**Planning**	**Architecture and design**	**Quality**	**Project tracking**	**Communication**	**Decision-making problems**	**Unforeseen ground conditions**	**Contractor' s efficiency**
(Calleam Consulting LTD, 2019)		X		X	X			X			X
Chan & Kumaraswamy (1997)		X	X		X				X	X	
(Moyce, 2014)	X										
(Marr, 2016)	X	X					X	X			
(Holgeid & Thompson, 2013)		X		X					X		
(Nguyen & Chileshe, 2014)		X		X	X						X
(Doloi, 2013)		X		X	X	X	X	X			
(Sitwala & Wium, 2013)			X	X							X
	2	6	2	5	4	1	2	3	2	1	3

2.6. Conclusion

The Project Management Institute (2016) lists lack of planning, poor pre-construction preparation, poor communication and teamwork skills, and weak contract administration as leading problem areas in construction. This correlates with the information in Table 2-10 where the various factors from the research studies are collated. The frequency of occurrence (AR) is calculated for each of the factors by dividing the times of occurrence by the number of case studies. The factors with the highest occurrences are leadership, competent project teams, planning, setting goals and objectives, and communication.

Table 2-10: Summary of factors from case studies

Factor	**Research Study Number**										**AR**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
Cost	X		X								2/10
Quality	X		X								2/10
Time management	X		X								2/10
Safety	X										1/10
Environmental performance	X										1/10
Project planning and review		X		X		X		X	X	X	6/10
Goals and objectives		X	X	X	X						4/10
Effective governance/leadership		X	X	X	X	X	X	X			7/10
Competent project teams		X	X	X	X				X		5/10
Communication (with stakeholders)			X	X	X	X		X			5/10
Experienced design team				X	X	X					3/10
Funding				X							1/10
Contractor's efficiency						X				X	2/10
Unforeseen ground conditions							X				1/10
Low speed of decision making involving all project teams							X				1/10
Client-initiated design changes							X			X	2/10
Necessary variations of works							X		X		2/10

The traditional success factors measured in terms of time, cost and quality are no longer enough to determine the success or failure of a project. Through the literature review, some of the softer factors/reasons such as the planning of a project and the day-to-day operations of a project have been shown to have a far greater effect on the outcome of the project than traditionally thought.

Being able to determine the possible outcome of a project early on during the execution phase could increase the likelihood that the outcome can be changed. Identifying and measuring the factors that influence the outcome of a project is the first step of this process. This study investigates a possible way of doing this by using the large amounts of information that are generated on construction projects and by data mining the information to find trends and patterns that can be used to train a prediction algorithm. This algorithm can then be used to identify the markers in existing or on-going projects to predict the outcome of the project. Following the Knowledge Discovery in Data (KDD) process, the algorithms can be updated continuously to increase the accuracy of the predictions as more and more quality data is fed into the system. This research will contribute to the growing knowledge base of construction project management and provide application of modern computer science and business process automation and optimisation to this field of management.

3. Literature Review: Knowledge Discovery in Data

3.1. Introduction

This chapter comprises a literature review of data mining and Knowledge Discovery in Data (KDD). Four case studies are discussed that provide an indication of the current status of KDD in the construction industry. The chapter also contains an introductory section about data mining (Section 3.2). This chapter satisfies Objective 2 of the study, which is to determine the status of data mining and KDD in the construction industry.

Choudry et al. (2011) describe KDD as the process where useful knowledge is discovered from data, and where data mining is a specific step in the multi-step process of KDD. The analysis of a database can be done without a hypothesis from the user and can answer questions that the user did not even know to ask. KDD can be used to make predictions about future data, based on using the current data (Zarsky, 2003).

Holgeid and Thompson (2013) say that merely listing critical success factors does not necessarily reveal relationships between the factors.

Soibelman and Kim (2002) indicate that large amounts of data are collected on construction projects and are eventually is discarded without processing the data to extract any useful information. This is echoed by Hammad and AbouRizk (2014). Soibelman and Kim (2002) indicate that this information is not stored or used properly in the construction industry because of limited time available to construction managers for the analysis of data and because of the complexity of the data analysis which, in most cases, is not an automated process. One method of extracting useful information is by the use of a data warehouse application program. In simple terms, a data warehouse is a type of database that collates information from various sources (data bases) and is used to analyse the information contained in the various data bases. A data warehouse allows the user to perform in-depth analysis of the information (Cardon, 2014). To enable true knowledge discovery, the development of a data warehouse is merely the first step in the process. The information contained in the data warehouse should be clean. Data preparation for the data warehouse is discussed in Section 3.3.2.

Cao, et al. (2002) indicate that project managers need timely analysis reports to respond to deviations during the execution of a project. These reports should be easily obtained by the construction or project manager through an easy-to-use interface. Simple query-based

analysis of data can be performed on the information stored in a data warehouse. The data query can be done on-line, and off-line where the queries have already been processed and the report is simply sent to the user.

This contrasts with the study done by Hammad and AbouRizk (2014) where they use Online Analytical Processing (OLAP) to generate reports and graphs. The advantage of OLAP is a decrease in space requirements for query storage. However, it takes longer, and processing power is required to run queries for OLAP applications. The OLAP application is discussed in more detail as part of the research study in Section 3.4. Considering the complexity of the mining algorithms and the size of the data being mined high performance computing is a key ingredient in time-sensitive situations.

3.2. Data Mining

Hammad and AbouRizk (2014) state that, during the data mining process, datasets are analysed to find unsuspected relationships. Data are summarised in novel ways that are understandable and useful. There are numerous methods of data mining. According to Alton (2017) and Botha (2018), the following data mining methods are the most important data mining techniques used. The list is not complete, nor is it comprehensive. Still it serves as an indication of available methods.

- **Similarity and distance functions** are used to determine the similarity between data points. This forms the foundation of most data mining models.
- **Prediction** is used to predict the future based on past and current trends. It is normally used with other data mining methods.
- **Association pattern mining** is used for finding patterns from data based on the probability of data points occurring together.
- **Cluster analysis** uses several similarity and distance algorithms to group similar data points together.
- **Outlier analysis** uses similar to cluster analysis except that the outliers that are not similar to other data points are determined.
- **Classification and regression** use algorithms to determine patterns in the data that can be used to predict the class or value of unknown or new data points. Neural networks fall under this category.

- **Text mining** is similar to text analytics and uses specialised techniques to create a structure or order from the unordered text that can be used in other models, such as classification.

The following four sections feature research studies that describe the current use of KDD and data mining in the construction industry, as well as some reasons why it is not being used.

3.3. Research Study 11: Using KDD in Construction Projects

This section focuses mainly on the research done by Soibelman and Kim (2002). The researchers set out to use KDD technologies to discover patterns in in a very large database that was previously thought to be chaotic. They used information stored in the Resident Management System of the U.S. Army Corps of Engineers (USACE) to test the feasibility of their KDD model. They acknowledge the fact that true knowledge cannot be obtained from a database where data has been collected and stored inconsistently. The researchers feel that construction data are usually limited in their breadth when it comes to using data to determine casual relationships related to schedule delay. For this reason, it is important to have in-depth and detailed data-preparation techniques. Data preparation can take up to 60% of the relevant effort in the KDD process, highlighting the importance of having correctly allocated and clean data. Figure 3-1 illustrates the KDD process used by Soibelman and Kim (2002) in their case study. It shows the detailed steps and the types of output that could be generated with the available tools. Each one of these steps will be discussed briefly based on the work by the authors.

3.3.1. Identifying Problems

Identifying the problem to be solved is a key step in engineering. To enable this, it is important to accurately identify and define the problem. Obtaining the domain knowledge for the relevant topic (in this case construction schedule delays) can be done through expert questioning and literature review.

3.3.2. Data Preparation

Data preparation is a particularly important step in the KDD process. The person analysing the data does not necessarily know how the data was collected. There may be missing fields, errors in the entries that were made or even ad-hoc data-collection policies

employed. The actual data mining effort normally represents only around 10% of the effort spent, while the analysis and assimilation of the data is around 10% of the effort required for knowledge discovery in databases (Soibelman & Kim, 2002). The proper management of data and data input can reduce the relative effort required for data preparation. Some of the problems that can be present in a database include missing parameter values, improper data types, out-of-range data, incomplete records or instances, and unavailable data (Soibelman & Kim, 2002).

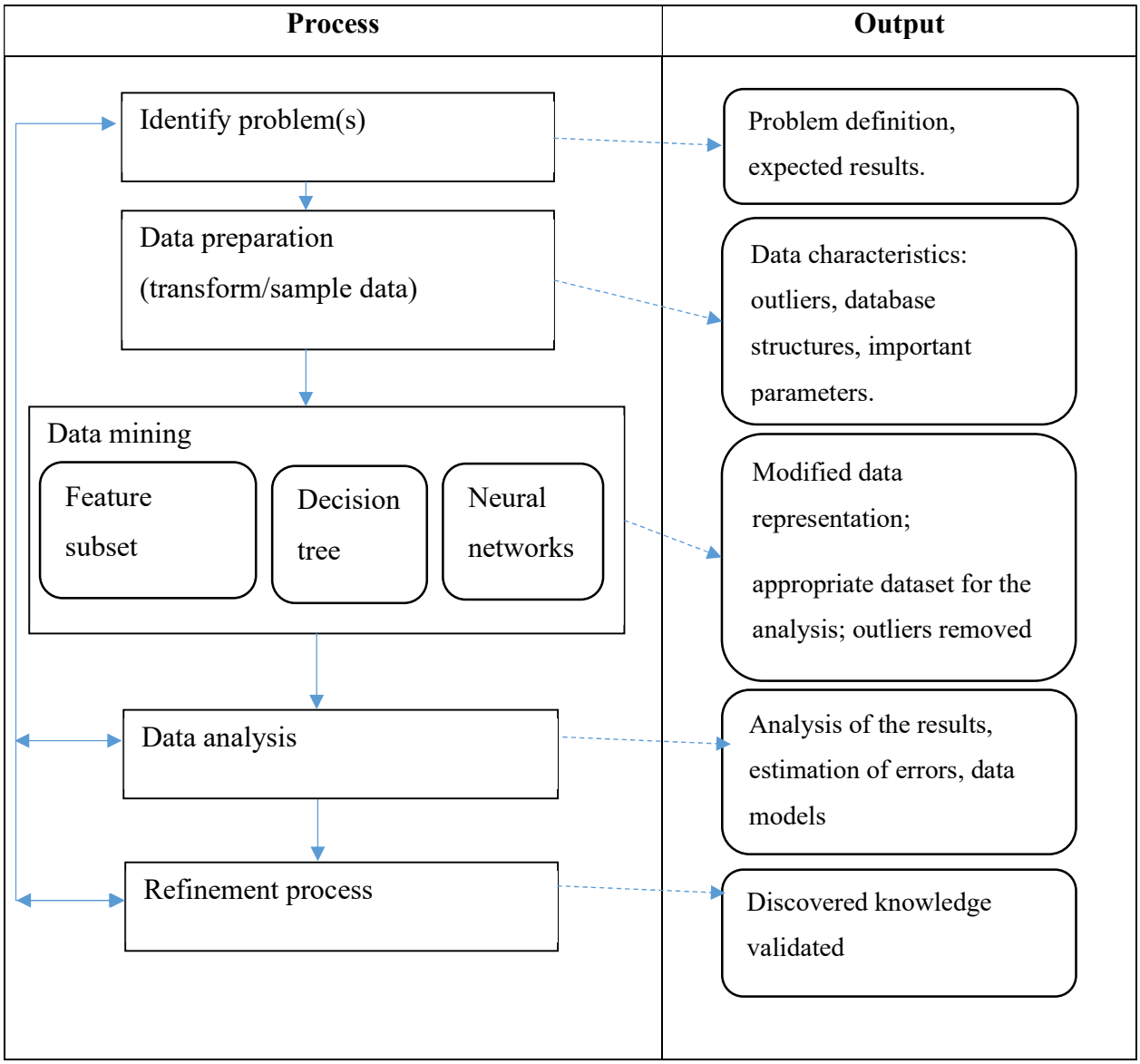


Figure 3-1: The main processes of discovery approach (Soibelman & Kim, 2002)

When preparing data for use in a KDD system, one needs to take four actions to prepare the data (Soibelman & Kim, 2002).

1. Removing variables that contain no to almost negligible information. Variables that contain information that is not relevant or of no use to the analyst should also be removed to reduce the total number of variables to a minimum.
2. Outliers should be checked and corrected where it has been established that they were captured incorrectly. The outliers should be discarded after confirming their status as outliers and, therefore not relevant to the data in question.
3. It is important to consider the variability in the data and the confidence level that the representative dataset is a true reflection of the data contained in the complete dataset when dividing a given dataset between different train-and-test sets. The confidence level of the accuracy of the dataset can be obtained by calculating the standard deviation of the full data and the test dataset.
4. All data mining tools can handle numerical data, but some software cannot handle alpha data. When converting alpha data to numerical data, it is important to not simply allocate arbitrary numbers to the alpha data.

3.3.3. Data Mining

Soibelman and Kim's (2002) research aims to find a data-analysis mechanism that can be applied to find patterns that explain or predict behaviours in construction projects for the U.S. Army Corps of Engineers. They start off by using feature-subset selection, then a decision tree and, ultimately feeding this into neural networks to make predictions of the future trends in a construction project.

3.3.4. Data Analysis

Soibelman and Kim (2002) use the results from their decision tree to create a set of rules for use in neural networks to predict trends. They use a back propagation neural network in their study. The network starts off with a random set of weights. The weights are adjusted after each learning cycle.

The learning must be repeated for each case in the training set by adjusting the weights after each output. For their specific study, they find that the best results are obtained by using a 1% learning rate and a three-layer back propagation neural network. Neural networks are explained in Section 6.4.2.1.

3.3.5. Refinement Process

During the refinement process there is a need to validate the results from the data mining process. Soibelman and Kim (2002) indicate that the average number of days (as retrieved from the study data base) for an activity of 320 units using 10 workers was 3.2 days. The results they obtained from the neural networks were between 4.96 and 6.86 for the same activity. This indicates that average information obtainable from the secondary database does not consider all the variables that can influence the actual duration. The writers continue their validation by running a Monte Carlo simulation with the same confidence level as the neural network. The Monte Carlo simulation gave a result of 3.26 to 15.5 days for the same activity. They indicate that, because the result is probabilistic in nature, the experience of the user is required to determine the actual answer to be used. Soibelman and Kim (2002) state that the neural networks provide the user with a more realistic answer. In their conclusion, they prove the consensus of the site managers that it is untrue that the weather was the dominant delay factor. The true dominant delay factor is equipment, and they indicate that a small, upfront investment in equipment could potentially have large cost savings for the project.

3.3.6. Summary

This case study is included because it relates to the use of information gathered during construction which is subsequently analysed using the KDD method to determine the reason for the late completion of projects. This enabled the authors to make recommendations to the management of the construction company so future projects can benefit from the new knowledge that was discovered.

3.4. Research Study 12: Knowledge Discovery in Data

Hammad and AbouRizk (2014) set out to use a five step KDD model to extract useful information from completed projects to make predictions about future projects. The authors use data for labour resources, generated while managing industrial construction projects from a large Western-Canadian structural steel fabricator, as their target dataset.

The various steps of the hybrid KDD model used by the authors are shown in Figure 3-2, and the five most important steps of the KDD model can be summarised as:

1. Understanding the problem domain;

2. Analysing the data and its multiple dimensions;
3. Developing the data collection model and prototype data warehouse;
4. Data mining; and
5. Evaluation of discovered knowledge.

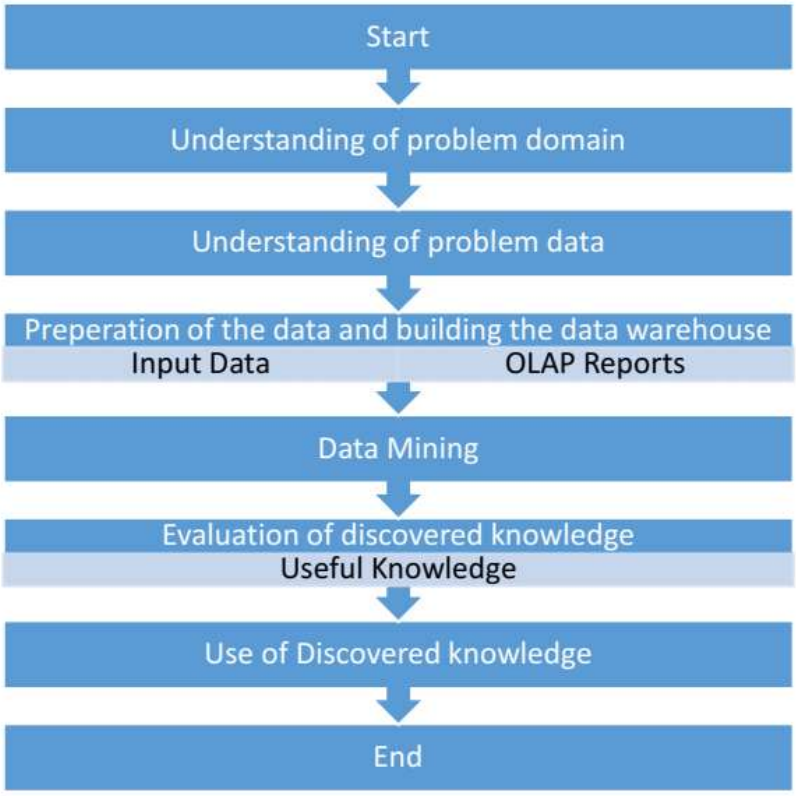


Figure 3-2: Modified hybrid KDD model (Hammad & AbouRizk, 2014)

The following sections provide a brief breakdown of the individual steps of the KDD process followed by the authors. Some of the information has already been discussed in Case Study 11 and only the information relevant to this case study is discussed under the specific heading. The case study uses three distinct datasets. Each dataset was handled on its own with the KDD process being applied to each dataset. The first dataset is used to prove that data mining can be used to improve and increase the efficiency of labour estimating practices used for tendering purposes. The second dataset is used to prove that mining of historical data can be used to better estimate the duration of work packages. The third dataset is used to develop probabilistic resource utilisation histograms for staffing of the project. The remainder of this case study is about the second dataset.

3.4.1. Understanding the Problem Domain

The current method to estimate the duration of activities is to divide the total man-hours for an activity by the number of working hours in a day. This can, however, be subjective and not provide a true or realistic estimate of duration based on current conditions. The dataset used contains actual durations and working hours for a large group of fabrication projects.

The dataset contains the following parameters: (1) start and finish dates for each activity, (2) actual weight of steel, and (3) working hours to complete each fabrication project. However, the project type was not assigned to the obtained dataset.

The purpose of applying the KDD process to the dataset is to use historical data to create representative and dependable estimating units that, once multiplied by the known quantities, provide estimates of the total duration and resource requirements of a project.

3.4.2. Data Cleaning and Pre-processing

Only information, where the start and end dates of the specific projects were marked as actual was used (in other words, completed projects). Any obvious mistakes, such as negative values, were removed from the dataset. The information for small packages was removed as these are manufactured by a different facility. Once complete, a large dataset of nearly 5600 data points was available to analyse. The original dataset contains nearly 13500 data points.

Prior to storing the data in the data warehouse, the following formula was used to determine the duration ($D_{(n)}$) in work weeks:

$$D_{(n)} = \frac{(Finish\ Date\ -\ Start\ Date)}{5}$$

3.4.3. Data Mining

The dataset contains actual data stored over a long period and had never been used for data mining or knowledge discovery prior to the case study. To verify the validity of the new knowledge, the dataset is split into two parts (15% and 85%). The larger set is used for data mining and knowledge discovery while the smaller set is used for testing purposes.

Clustering, was used to analyse the data. Each cluster analysis was repeated three times to ensure the stability of the results.

The following results are obtained from the clustering technique:

- Nineteen clusters are obtained for the hourly unit cost of fabrication;

- Thirteen clusters are obtained for the weekly unit duration of fabrication;
- Five clusters are obtained for the shop drawings' hourly unit cost; and
- Six clusters are obtained for the shop drawings' weekly duration.

3.4.4. Analysis of Discovered Knowledge

The clusters with the highest number of data points are most likely to represent the common types of work in the company, while the clusters with the lowest number of data points most likely represent rare types of work or outliers that have to be further investigated.

The second part of the dataset is used to validate their analysis. The results obtained from the analysis for the unit costs and durations are used to estimate the resources required and the duration of each of the projects in the remaining dataset (15%). The error percentages for each project are also calculated. The means of the two clusters are used to estimate the total resource requirement and duration for each project.

The results from the validation test shows that more than 80% of the dataset has an estimation error below 25%. Therefore, proving a significant increase in the accuracy of estimating practices when relying on historical data.

3.4.5. Summary

The result from the authors' analysis supports their claim that data can be transferred into useful knowledge that can provide the users with meaningful and new insights. This case study is included to show what information can be learnt through the KDD. In this case, study Hammad and AbouRizk (2014) prove that, by using the KDD process, more accurate information can be obtained for tendering purposes. They also show that, with enough information, predictions can be made based on historic data. This relates to what this study aims achieve.

3.5. Research Study 13: Predicting Project Performance in the Construction Industry

Assaad et al. (2020) set out to (1) quantify the impact of risks related to the performance of a project by incorporating a broad spectrum of inputs; (2) formulate a holistic assessment model; and (3) correlate the developed system to predict cost and time at project completion.

The authors commence with a review of previous studies and of the limitations of current predictive models. Previous studies that attempt to predict project performance by using various modelling techniques are mentioned. Assaad et al. (2020) conclude that most of the research focuses on schedule or cost as aspects of project performance, while others fall short of covering the varied risks that are present in construction projects. The authors argue that no research work offers an integrated approach to estimate the performance of construction projects. The authors mention different techniques to measure project performance:

- Multiple linear regression for construction projects in China using project management practices that were adopted by foreign architecture, engineering and construction firms.
- Discriminant function analysis method based on six practices: pre-project planning, constructability, project change management, design/information technology, team building, and zero-accident techniques.
- Statistical analysis and artificial neural networks and an identified list of critical factors that affect the performance of a project.
- A probabilistic method to forecast costs using Bayesian inference and the Bayesian model average technique.
- Risk integration and Markov-chain simulation to estimate the cost at completion of projects.
- Various other techniques or variations of the above-mentioned techniques to develop models for predicting deviations in project schedules and future progress and predictions of future project costs.

The 25 project risks that Assaad et al. (2020) identify as having an impact on project performance are based on a literature study spanning from 1985 to 2018 (Figure 3-3). The authors consider factors that overlap with those mentioned in Chapter 2.

Assaad et al. (2020) developed a survey to establish the probability of occurrence and impact of each of the identified project risks. This enables them to calculate the criticality to fit distribution functions for each of the project risks. A five-point Likert scale is used to minimise the respondents' biases.

Assaad et al. (2020) normalise the data so that they can fit into a scale of [0%, 100%] in order to fit the criticality that was calculated. The empirical distributions for the normalised data collected by the authors are shown in Figure 3-3. The normalised data allow the authors to fit

parametric distributions. The various project risks are fitted with different distributions that reflect their distinct impacts on the performance of the project. The factors that cannot be fitted with parametric distributions share common characteristics, such as complexity in nature and being able to quantify them after the fact.

Assaad et al. (2020) use the distributions to calculate cost and schedule overruns. They find that fitting triangular distributions to the data yields more accurate representations of the datasets. During this step, they also calculate the project risk weights which are a secondary indicator of the impact of each of the project risks.

Assaad et al. (2020) find that their model is more suited to industrial projects. This is because they obtain the data they use for the schedule and cost overruns from the Construction Industry Institute's COAA major projects benchmarking summary report. The project weights should also be adjusted to be project specific as projects are bound to unfold differently.

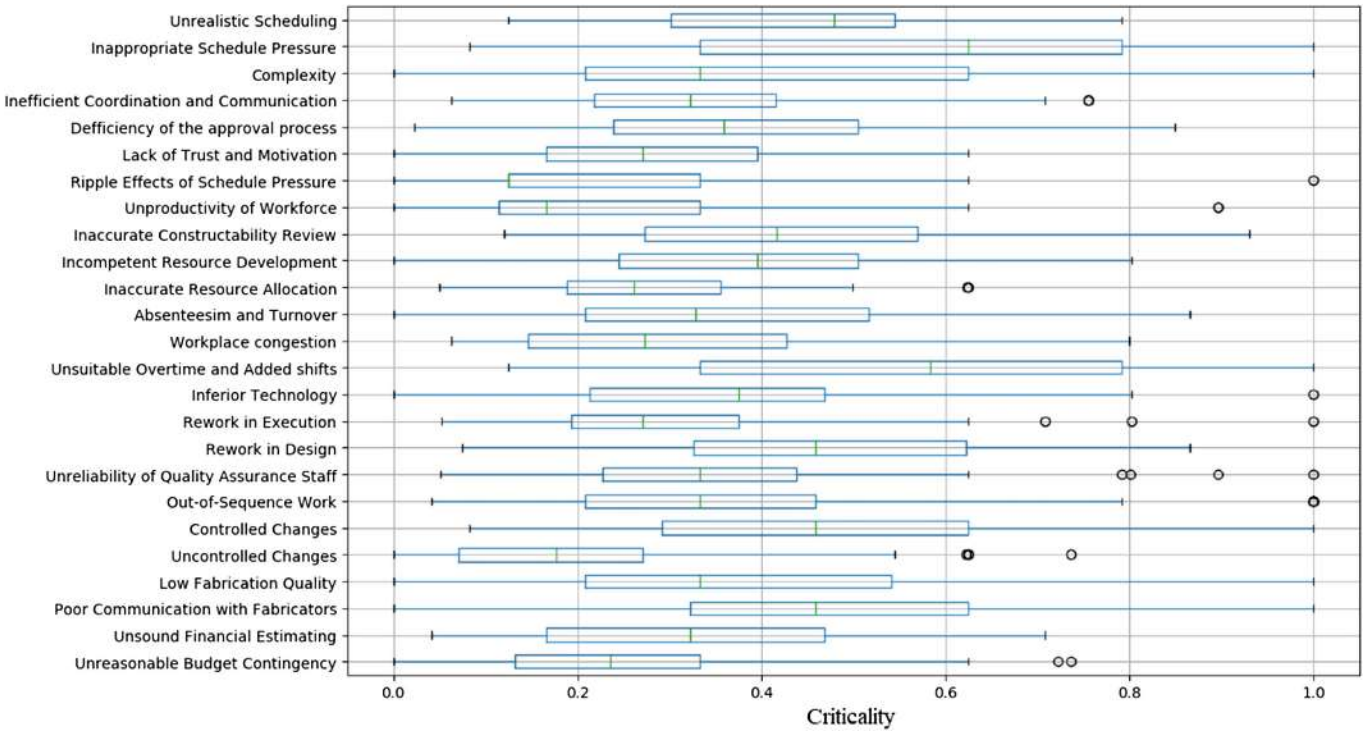


Figure 3-3: Normalised collected data for criticality (Assaad, et al., 2020)

3.5.1. Summary

Although this case study is performed using major projects in the USA, it is relevant to this research as it highlights what can be learnt from new methodologies in examining existing construction project datasets. Assaad et al. (2020) do not use data mining techniques or KDD,

but simple distribution fittings to create a model for predicting project performance. This demonstrates the complexity involved in trying to use a wide variety of factors during the KDD process. The main reason researchers focus on a single aspect to develop a model is due to complexity and time constraints. As indicated in Chapter 2, there is an overwhelming number of factors that contribute towards the success or failure of a project. To develop a model that takes all the aspects into account would be impossible. The permutations and combinations of factors that could influence the outcome of the project are infinite.

3.6. Research Study 14: Factors Affecting the Utilisation of Big Data in Construction Projects

Yu et al. (2020) aim to find factors that meaningfully impact the use of big data and how these factors influence each other. The authors indicate that a knowledge gap exists limiting the promotion of big data utilisation in construction projects. This is because no comprehensive list of factors, nor their interdependencies on each other, exists or has been fully investigated.

Through a literature review and through interviews with experts, the authors compiled a comprehensive list of factors that can influence the use of big data in construction projects. According to the authors, the characteristics of big data in construction projects can be summarised as six V's:

- The **Volume** of data can be massive and complex. The example the authors use is BIM which, for a design of a three-storey building, can easily reach 50GB.
- The **Velocity** of the data is the speed at which the data is produced and processed on construction sites.
- A **Variety** of data formats are used to store data on construction projects. This could be DWG for the drawings, JPEG for pictures taken during construction and various other formats that can be used to store the generated information.
- The **Veracity** of the data provides an indication of the quality of the data. Noise and inaccurate attributes frequently emerge from large volumes of data.
- The **Variability** of the data flow rates take place during the life cycle of the project. The authors use the following example: A project manager may receive a large

amount of BIM data at the start of the project when the designs are completed but may not receive any additional BIM data for an extended period during construction.

- **Value** is an indication that big data should generate additional value to the project stakeholders and the project owners.

Yu et al. (2020) highlight some areas where big data and related technologies have been adopted and implemented in construction projects. Some of the areas include stakeholder collaboration, cost-, safety-, tendering-, waste-, equipment- and supply chain management.

According to Yu et al. (2020), their literature study reveals the following factors that affect the utilisation of big data in construction projects:

1. Availability of big data technologies;
2. Organisation structure and culture;
3. Stakeholder management and collaboration;
4. Availability of big data facilities;
5. Standards, guides or handbooks for utilising big data;
6. Ethics including copyright, confidentiality and data security;
7. Client satisfaction;
8. Post-project evaluation;
9. Incentive policies;
10. Effect of big data on project performance;
11. Robustness and effectiveness of big data;
12. Exemplary projects;
13. Top management support;
14. Human resources;
15. Data sharing and governance;
16. Conservatism; and
17. Disjointed nature of the construction industry.

Following the literature review and expert interviews, the authors model the interrelationships of the identified factors by using interpretive structural modelling. Using expert calibration, the results are calibrated and verified after four rounds of open discussions between the authors and the experts.

The factors are classified into several clusters to determine the driving power of the individual factors. The authors conclude that Factor 6 (copyright, confidentiality and data

security) and Factor 9 (incentive policies) have the highest impact on the utilisation of big data in construction projects. The authors argue that these factors could be considered as external environmental elements and, to increase the utilisation of big data, more attention should be paid to the ethical and legal issues and incentive policies for big data utilisation.

The next level of criticality that has been identified by the authors includes Factors 1 (availability of big data technologies), 13 (top management support and big data-orientated management modes), 14 (human resources), 15 (data sharing) and 17 (governance). The authors argue that the results show that management practices of construction projects should be modified to reflect the big data orientated operations. This means that stakeholders and ordinary employees should actively engage in the decision making for the project to improve data sharing and knowledge exchange.

3.6.1. Summary

This case study highlights the importance of big data in construction projects, and emphasises the need for South African construction and engineering companies to use this technology to become leaner and more competitive, especially during difficult market times.

The characteristics of big data as highlighted by the six V's, assist in ensuring that any information being mined contributes and add value to the project and stakeholders.

The factors affecting the utilisation of big data are relevant as there needs to be a buy-in from senior management to implement the system. Once the value from using big data to predict project outcomes is realised, it will be implemented and used widely.

3.7. Discussion and Conclusion

Objective 2 of this research is to determine, from literature, the current application of data mining and KDD in the construction industry. The four research studies discussed in this chapter mention some of the uses of KDD in the construction industry and the reasons why it might not be fully implemented in the industry. No case studies could be found where data mining was applied to the South African construction industry.

Objective 4 of this research requires that the information obtained from project meeting minutes be analysed through the KDD process. It is, therefore, vital and relevant to note the important steps that are present in all the KDD case studies presented. These steps are goal

definition, data preparation, data mining, data analysis and refinement (which includes model verification and/or validation).

From the case studies, the importance of having a clear and achievable goal is evident. This will inform the investigation of the data to be collected, how the output data should be formatted, and how the results will be utilised. It will also inform the investigation of the data mining method and accuracy required.

After defining the goal, the next important steps in data mining are data acquisition, assimilation and preparation. The user should understand where the data comes from and what is captured for them to make meaningful decisions on what features to use. This will be used to correct missing and incorrect values and determine what data mining process should be used to provide the best results.

Finally, the user should perform data analysis, which includes validation and evaluation of the data mining output, from where the process can be refined to obtain better results. The data analyst must understand the business, know what to model and how to apply it.

4. Research Methodology and Data Collection

4.1. Introduction

This chapter describes the research methodology adopted for the investigation of the contents of several project meeting minutes discussed in Chapter 6. The factors identified from the literature review were transformed into features and are used to create the datasets for data mining. The investigation illustrates the benefits from data mining existing documents – in this case, the meeting minutes of a project.

4.2. Research Methodology

From the literature review, two overarching categories contributing to the outcome of a project are discussed in Chapter 5. From this, several project attributes are identified that are used in the data mining investigation.

The project attributes identified are used in an investigation of several project meeting minutes in Chapter 6. There, the site meeting minutes from past projects are used in two different data mining applications to verify the impact of each attribute on the outcome of a project.

4.3. Data Collection

Meeting minutes are used from 27 civil engineering projects that were constructed in South Africa over the last 10 years. The meeting minutes were provided by a construction company who undertook most of these projects. The types of projects range from heavy civil engineering to marine, infrastructure and rehabilitation projects. There are several repeating clients due to the nature of the projects (such as road construction and marine projects). The clients and contractors are not part of the factors being examined and have been left out of the datasets that were created.

4.4. Data Processing

The initial dataset (called the primary dataset), created from the various project meeting minutes, was used as a baseline to create an additional three datasets. The new datasets are

exact copies of the original one, except that they have been edited to keep only the relevant information as explained below.

One of these datasets was edited to keep a select number of meeting minutes. The second dataset was edited to leave just text (no numerical values). The third dataset was edited to fill all the missing information with an arbitrary value (0). The workflow of creating the datasets is shown in Figure 4-1 with the four distinct datasets. The selected meeting minutes dataset (Dataset S) contains 12 projects with a total of 209 individual meeting minutes. This dataset was created to ensure that all the features are at least 90% complete.

The rationale behind creating the different datasets in addition to the primary dataset is to determine (1) if the amount of missing or incomplete information affects the accuracy of the prediction models created; and (2) if the text-only dataset is required for text data mining.

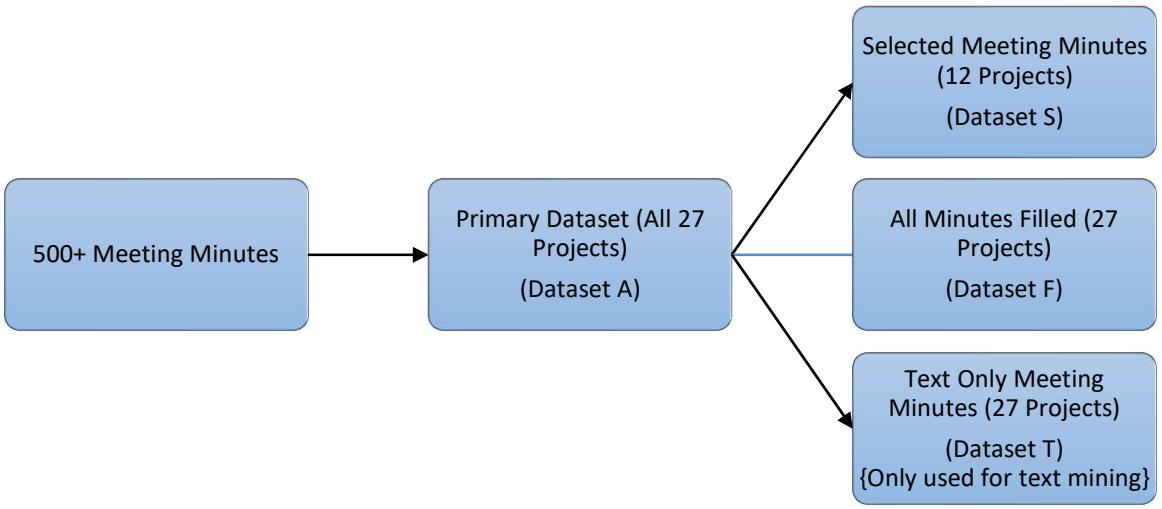


Figure 4-1: Workflow for creation of datasets used in data mining case study

4.5. Data Mining

Two different data mining applications are used to determine if there are any discernible differences in the type of application used but, also to verify the results from the applications against each other. The same datasets are used for both applications to allow a comparison.

The two applications have been chosen based on their relative ease of use, availability and hardware requirements. The applications used are Orange and RapidMiner. Both applications

are dedicated data mining applications with visual inputs. The two data mining applications have been chosen above others such as Weka and MATLAB as the latter applications require extensive knowledge of the different programming languages that they use.

Various algorithms are applied to the different datasets to determine the optimum prediction models that could be created with the given information. Each application has its own set of built-in algorithms that can be used to create the prediction models. The algorithms that are used are summarised in Table 4-1. Most of the algorithms are the same for both applications. Orange has a k-Nearest Neighbour algorithm that was not used in RapidMiner, while RapidMiner has a fast large margin algorithm not used in Orange.

Table 4-1: Algorithms used in data mining case study

Orange	**RapidMiner**
k-Nearest Neighbour (kNN)	-
Logistic regression	Logistic regression
Linear regression	Generalized linear model
Decision trees	Gradient boosted trees
Random forest	Random forest
Neural networks	Deep learning
Support Vector Machine (SVM)	Support Vector Machine (SVM)
Naïve Bayes	Naïve Bayes
-	Fast large margin

5. Analysis of the Most Relevant Success Categories

5.1. Introduction

There is rarely a single factor responsible for the success or failure of a construction project. Under normal circumstances, it takes a confluence of various factors. This chapter focuses on the two main categories identified in Chapter 2 and how the various other factors can be grouped together as contributing factors under the two main categories. Section 3.5 of the literature review provides a list of important attributes to use during the KDD process.

In the literature review in Chapter 2, a few of the factors are cited more than others. These factors are leadership, competent project teams, planning, setting goals and objectives, and communication. For the purposes of this study, leadership and competent project teams are considered to be related factors. Setting goals and objectives is a critical skill of a good leader, but this also influences the planning of a project. If no goals or objectives are defined at the start of a project, and then monitored and adjusted as required during the execution of the project, it might not be possible to track progress on the project. Communication plays an important role in a project team, and it is important to have as a good leader. Not all the factors can be unpacked and individually addressed. Thus, some of the factors have been combined into two broad categories that will be used in this study. The two categories that are used are leadership and planning. They are discussed below.

The project attributes described below are used as features in the data mining process to determine if they influence the outcome of a project.

5.2. Leadership

Leadership is a broad term made up of a powerful combination of cumulative knowledge from different disciplines, such as anthropology, political sciences, psychology, sociology, education and business. (Stjepanovic, et al., 2020). According to Windapo et al. (2015), leadership is provided by the project manager by establishing direction for the team and motivating and inspiring subordinates.

The Association for Project Management's Book of Knowledge (2012) defines leadership as:

> *"The ability to establish vision and direction, to influence and align others towards a common purpose, and to empower and inspire people to achieve*

> *success. It enables the project to proceed in an environment of change and uncertainty."*

In their fourth edition of the PMBOK (2008), the Project Management Institute defines leadership as:

> *"The ability to guide the project team while achieving project objectives and balancing the project constraints."*

The International Project Management Association (2020) defines leadership as follows:

> *"Leadership is the ability to establish vision and direction, to influence and align others towards a common purpose, and to empower and inspire people to achieve success."*

According to Zahra et al. (2014), a project manager must have skills in three major dimensions – technical, transactional and transformational. It is in the transformational domain where the leadership activities are categorised. They include communication skills, people skills, negotiation skills and conflict-resolution skills.

Windapo and Akintona (2015) conclude that project customer satisfaction tends to increase as the general skill of project managers increase. It is, therefore, important to ascertain – as far as possible – the level of leadership that is involved in the project.

People can experience leadership in different ways. Stjepanovic (2020) indicates that, if team members show a change in their behaviour while executing their responsibilities, one can say that they have been influenced by the leader.

For this study, the effectiveness of leadership will be measured through the following attributes in the site meeting minutes of the project:

- Setting goals at the start of the project;
- Regular meetings where progress is discussed, and feedback is given to the stakeholders;
- The number of attendees present at the meeting; and
- The variance in the anticipated completion date (between subsequent meetings).

Furthermore, a large portion of the costs of a construction project can be attributed to rework. Burati et al. (1992) indicates that from their research, an average of 17% of the total project cost can be attributed to rework. They indicate that design deviations contribute to an average

of 9.5% of the total project cost. This research also includes the possible effect of rework and redesigns on the outcome of a project.

5.3. Planning

Effective pre-project planning and regular updates to the programme have been highlighted by several sources as a key component to project success. From their interviews, Gaetsewe et al. (2015) establish that a cause of time delays in construction projects in the Northern Cape in South Africa is a lack of understanding of construction projects. The authors mention that this lack of understanding is because of poor planning by the client who provides an unclear scope and design changes. This is further exacerbated by poor design co-ordination by the consultants. According to the authors' interviewees, poor planning during the design stage causes, amongst other things, changes in the scope of works, variation orders and the need to accelerate projects. They also point out that project time overruns and cost overruns are worsened by rework due to poor workmanship and delays resulting in the extension of time claims.

Zahra et al. (2014) point out that a project manager should possess the skills to plan, organise, control and monitor the project by doing budget and schedule planning with the end goal of making the project a success.

In the paper presented by Serrador (2012), it is concluded that there is a correlation between planning and the success of the project. When comparing the correlations with planning, he finds consistent empirical results indicating an average R^2=0.33 correlation with efficiency and an R^2=0.34 for overall project success. He argues that this indicates a significant impact on project success when compared to the reported 20% to 34% effort spent on planning. The author concludes that planning is associated with project success, and that the amount of planning at the start of the project is related to project success.

This study investigates, through the following items, how planning can be assessed from site meeting minutes:

- Considering actual and planned progress percentages together with the text commentary or notes about the progress being achieved on site, which could include programme revisions and updates.

- Having regular meetings where the project progress is discussed. During these meetings, the achievements, difficulties and frustrations, and the possible risks that might hamper the project should be discussed. The meetings should take place regularly as agreed to by the different parties. These will be different on each project but, normally, they will take place monthly, fortnightly or even weekly if the project so requires. The time between meetings might be a factor that could influence the outcome of a project.
- Tracking the milestone dates and the progress towards them. This could be done by observing the anticipated completion dates and by comparing the variances between the subsequent meetings. This may indicate inconsistent planning and leadership.
- Confirming and recording project delays are also important. This could stem from a lack of planning. attributes that can be recorded under the heading "delays" can include the number of claims for Extension of Time (EoT) or whether there is a lack of information for certain areas.

5.4. Discussion and Conclusion

From the literature review and the discussion on narrowing down the factors to two broad categories, the following project attributes are identified as having an impact on the outcome of a project. These attributes were used as features for the data mining process in Chapter 6 if they are present in the meeting minutes:

1. Setting goals – what is the project about and what are the main priorities and outcomes for the project. This is not present in the meeting minutes and thus was not used.
2. Time between site meetings – This attribute is determined from the meeting dates.
3. Number of attendants at the meeting, both from the client's side as well as the contractor's side. Only having a few of attendees at the meeting might point towards bad leadership, while having too many attendees may indicate an unnecessary level of bureaucracy in the system.
4. Number of Requests for Information (RFI's) and Site Instructions (SI's).
5. Quality comments – especially if concerns are raised about rework and elements of a quality nature that are not being addressed by the contractor.
6. Decision making and problem resolutions – In other words, address outstanding actions on meeting minutes.

7. Progress updates from the programme (indicating the percentages of planned vs. actual) – The updates should also indicate how many days the project is behind or ahead of programme.
8. Number of Compensation Events (CE's) or claims from the contractor – This indicates the number of unforeseen problems that the contractor might have encountered. Although the claims themselves might not be approved, the mere fact that the contractor has submitted a claim should indicate that the project is not progressing as planned.
9. In cases where the NEC contract documents (or similar) are used, the number of early warnings that has been submitted by both the contractor and the client. The early warnings are preventative and should assist the project management team in preventing delays and claims. Therefore, it could prove to be an important factor in determining the impact on project completion.
10. Project completion dates and progress towards completion dates – confirming how much the dates have changed from the start of the project. Expected completion dates vs. the milestone dates might also show some interesting information, especially if they do not correlate with the percentages behind or ahead.

6. Investigation of Several Project Meeting Minutes

6.1. Introduction

This chapter presents an investigation of several project meeting minutes, as input to data mining. A knowledge discovery in data (KDD) process was applied to the meeting minutes to investigate if the project attributes mentioned in Chapter 5 can be used as features of a dataset to predict the outcome of a project. The predictions of the outcome of the project can be done at any time during the project, using the previous site meeting minutes already recorded for the project.

A typical KDD process involves five key steps. These five steps are presented in Section 3.3 and, in this chapter, are applied to site meeting minutes to satisfy Objectives 3 and 4 of this study. Objective 3 of the study is to determine if the project attributes identified in Chapter 5 are being recorded in the minutes of the project. Objective 4 of the study is to apply the KDD process as an example to determine what information would be required from the minutes to predict the outcome of a project.

The five steps of the KDD process are:

1. Problem identification and definition – For this study, the goal of the process is to determine if the attributes being recorded in the site meeting minutes predict the outcome of the project (Section 6.2).
2. Data preparation for data mining – The datasets created from the site meeting minutes and the features of the datasets are discussed (Section 6.3).
3. Data mining – The application of data mining algorithms from two different data mining applications. Each of the data mining applications is discussed separately with an overview and explanation of the data mining algorithms used (Section 6.4).
4. Data analysis, including validation – The results from data mining for each of the two data mining applications are discussed separately (Section 6.5).
5. Refinement process (Section 6.6).

Each of these steps is discussed under its own heading below.

6.2. Problem Identification and Definition

The first step of the KDD process is to identify the problem that needs to be investigated through data mining and analysis. The goal of this study is to determine if the attributes currently being recorded in the site meeting minutes predict the outcome of a construction project.

Through the process of data mining, the information obtained from the meeting minutes will assist in creating various prediction models that can be used to predict the outcome of projects. Secondly, it will indicate whether any new knowledge can be obtained from the meeting minutes through the KDD process. True knowledge cannot be obtained from the meeting minutes in their current formats. The meeting minutes do not follow a sequence (or format) that can easily be used.

Therefore, the problem identification or goal definition can be phrased as: By using previous project data stored in a database, can the information from project meeting minutes be used to predict the outcome of a project?

6.3. Data Preparation

This section discusses the creation of the data warehouse, the dataset features and the pre-processing of the data.

The meeting minutes from a total of 27 civil engineering construction projects were obtained for this study. Among these 27 projects, there are more than 500 sets of individual site meeting minutes. The projects were all completed in South Africa and range from repair or maintenance type projects to new construction projects in the infrastructure, heavy civil and marine engineering environments.

Due to the nature of the projects, several were completed for the same clients in different parts of the country. The project teams from the client and the contractor were different in these cases and, thus rule out familiarity and bias in the running of the project.

The format of the minutes differs drastically from one project to another. Even when projects were completed for the same client and the basic formats were the same, the level of recorded information differed. Most of the meeting minutes were stored in PDF format while some of the minutes were stored in the native MS Word format. Some of the PDF documents were

scanned copies, which could not be used directly by a computer without review by a data editor after the text recognition process.

Since the meeting minutes are all recorded in table format with text and lists, the data can be described as semi-structured. It means that it is easily understood by a human reader but cannot typically be used by a computer system.

The site meeting minutes can be further described as dependency-orientated data, meaning that the data contains attributes that have internal relationships between the entries. These dependencies are time-series data due to the continuous recording of the site meeting minutes over time. The duration of the contract, number of site instructions (SI's) and field engineering queries (FEQ's) are examples of the type of internal relationships that were encountered.

When this information is loaded directly into a data mining application without any pre-processing the computer software is only able to interpret some of the information. The Orange import widget was tested by loading nine PDF files from a single project – all with the same document layout -directly from the source folder. The test was done to verify whether the data mining applications could read the documents without any human intervention.

Figure 6-1shows an Orange screenshot indicating the information that the application was able to read from PDF documents. Figure 6-1-a shows a document where nothing could be read, while Figure 6-1-b shows a document where the application could read the contents. The information is shown as a continuous text document and can be used only as text. The data mining applications do not understand to which feature the specific information should be allocated. Installing an extension for RapidMiner provides a similar ability to read text from a PDF for text mining.

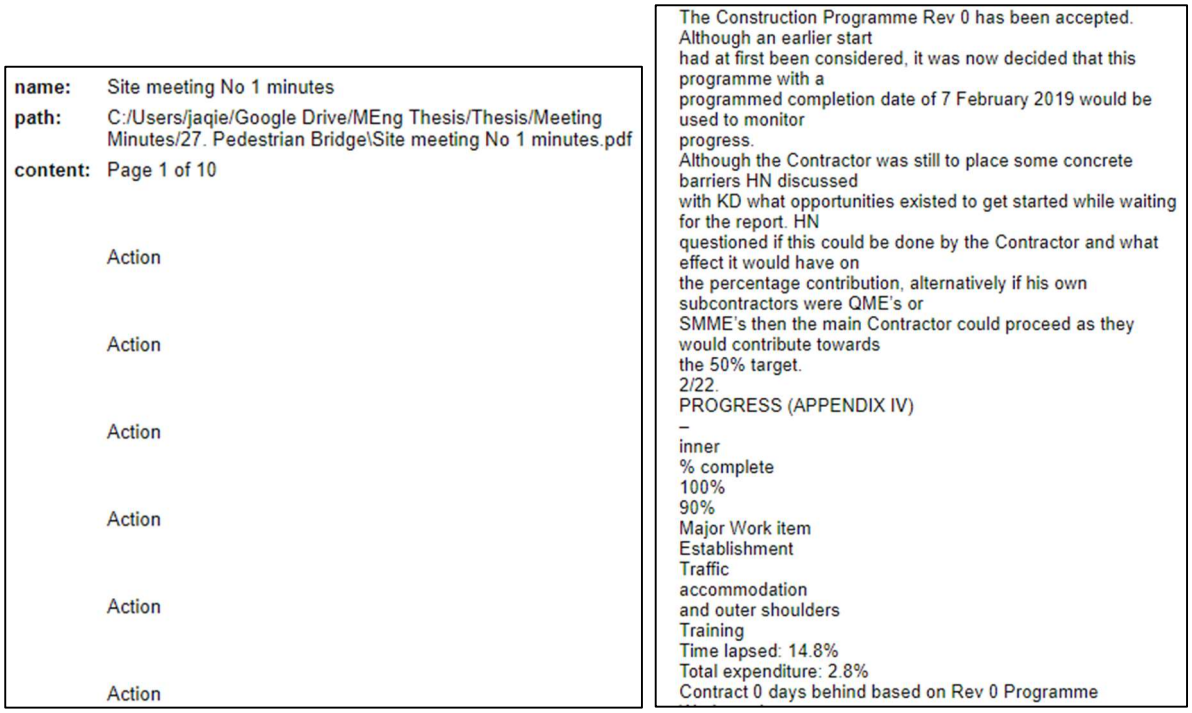


Figure 6-1-a (Left) and Figure 6-1-b (Right)

Figure 6-1: Screenshot of Orange document viewer comparing the contents of two documents

6.3.1. Creating the Data Warehouse

For this case study, the data warehouse contained only the relevant information from the meeting minutes. Normally, the data warehouse would contain the actual files (such as PDF and MS Word) with all the information. Because of the limitations mentioned earlier about the format of some of the information in the minutes, it was necessary to manually populate the data warehouse with the information from the meeting minutes. The meeting minutes of each of the 27 projects were transferred manually into separate spread sheets for each project. However, some projects did not have all the meeting minutes available.

The various attributes that were recorded together, along with a short description of each attribute, are shown in Table 6-1. The dataset contained 509 rows of information with 39 attributes used as features for data mining. The features were determined from the attributes described in Chapter 5. For most cases, the attributes were the same as features. There were instances, however, where some calculations or modifications were needed so the attributes could become features for use in the data mining applications.

Table 6-1: Attributes recorded in the data warehouse

Attribute name	**Type of information stored**
Meeting number	Continuous numbers representing the progressive meeting number of site meeting (in other words, 1, 2, 3…) as recorded.
Date	Date of the site meeting.
Number of contractor employees at the meeting	Actual number of contractor employees mentioned on the attendance register.
Number of client representatives at the meeting	Actual number of client representatives, or appointed engineer and representatives, on the attendance register.
Health, safety, environment and community	Text from health, safety, environment and community discussions recorded in the minutes.
Security and IR	Text from discussion related to security and Industrial Relations (IR) recorded in the minutes.
Quality	Text from quality-related discussions recorded in the minutes.
Contractual/commercial	Text from contractual or commercial discussions recorded in the minutes. For the purposes of this study, any discussions about the finances of the project were ignored. Exceptions were made in a few instances where the client's late payment resulted in project delays or where the contractor's financial difficulties resulted in the late completion of a project.
% Progress	Actual percentage progress from the updated programme. This can vary if the programme has been re-baselined due to changes on site (including compensation events).
% Progress planned	Percentage progress planned when the programme was updated. In cases where this was not available, but percentage progress was, it was calculated as the straight-line progress from the start of the project.
Expected completion date	The expected completion date of the project as indicated from the updated programme.

Feature name	Type of information stored
Progress notes	Any text that was recorded in the meeting minutes about the progress being achieved on the project.
Number of FEQ's	The cumulative number of Field Engineering Queries (FEQ's) raised by the contractor.
Number of SI's	The cumulative number of site instructions issued to the contractor.
Original/contractual completion date	The project completion date. This date can change as the project progresses and the contractor is awarded an extension of time.
Early warning from contractor	The cumulative number of early warnings from the contractor. This is relevant to projects that use the NEC documents as the contract documents of the project.
Early warning from client	The cumulative number of early warnings from the client. This is relevant to projects that use the NEC documents as the project's contract documents.
Number of compensation events	The cumulative number of compensation events (contractual claims for additional money and/or additional time) claimed by the contractor.
Days ahead or behind programme	The number of days ahead or behind on the programme at the time of the update.
Class of progress	If the meeting minutes mentioned that the project was behind, on time or late, it was recorded as such – irrespective of the days ahead or behind. This is because a project could be X days behind and still be on schedule to complete on time.
Project man-hours	The cumulative number of man-hours recorded on the project.
Weather delays	The cumulative number of weather delays experienced as recorded in the meeting minutes.
Rework mentioned	Mention of any indication that the quality of work was not up to standard and that the work had to be repaired or rebuilt.

Feature name	Type of information stored
Repeat items on the minutes	A Yes/No response was recorded for any important items that were carried over from one meeting to another without being addressed.
Design delays	A yes/No response was recorded for any mention of delays because of late design information or drawings.
Planning	A Yes/No response was recorded for whether the programme was being discussed or mentioned in the meeting minutes. For this to be deemed Yes, the programme must be discussed. Progress on structures or sections alone cannot be related to the planning of the project when considering the use of the minutes for data mining.
Project complete late	A Yes/No response was recorded for whether the project finished on time or late (taking into consideration any extension of time that was given).

The following project attributes were calculated from the information obtained in the minutes:

- Duration of contract elapsed (meeting date minus contract start date – measured in days);
- Progress variance (actual progress minus planned progress – expressed as a percentage);
- Days between meetings (meeting date minus previous meeting date); and
- Variance in anticipated completion dates (anticipated completion date indicated in the meeting minus the anticipated completion date of the previous meeting – measured in days).

Because all the projects were unique and could not be directly compared due to the different scales within the projects, the following attributes were normalised so that they would not skew the results:

- Meeting number;
- Number of contractor employees at the meeting;
- Number of client employees at the meeting (including engineering staff)

- Duration of contract elapsed; and
- Number of FEQ's and SI's.

6.3.2. Dataset Features

A useful feature – in both Orange and RapidMiner – is feature statistics. The feature statistics provide the user with information on the dataset, including the size, number and types of features that are present in the dataset (not shown in the example). An example of the feature statistics created by Orange is shown in Figure 6-2. In the figure, the feature with the largest amount of missing information is weather delays (56%). The minimum, maximum, centre and dispersion of the distributions are shown. This assists the data analyst in determining whether a feature should be used or ignored in the data mining process.

Appendix 3 contains the full screenshot of the information in Figure 6-2. It shows the statistical distributions of the features and a description for each of the columns of the widget. The appendix also contains an extract of the feature statistics from RapidMiner.

Name	Center	Dispersion	Min.	Max.	Missing
Design Delays	No	0.419			1 (0%)
Finish Late? (Yes/ No/Unsure)	Yes	0.692			91 (17%)
Class (Ahead/ Behind/OnTime)	Behind	0.656			106 (20%)
Days ahead/ behind	-23.1142	-1.70915	-264	0	122 (23%)
Weather delays (Rain;Wind etc)	7.7867	1.3827	0.00	46.50	287 (56%)

Figure 6-2: Orange feature statistics for selected meeting minutes

6.3.3. Data Pre-processing

In some of the meeting minutes, some information was not recorded, which makes the data incomplete. In certain instances, it was possible to obtain the missing information from the site meeting minute appendices (where they form part of the recorded site meeting minutes). In cases where individual values were missing, the missing information was appended by averaging the values of that specific feature. Where the missing information was text, the words "nothing to report" were inserted into the specific cell. This phrase can be ignored when the text is analysed. This was done by including it on a list of excluded words during the text pre-processing step during data mining. This addition of missing information was to make the dataset as complete as possible.

Some of the features adversely affected the outcome of the data mining due to the large amount of incomplete and missing information. These features were left out of the final data mining analysis. The features that were left out are:

- Early warnings from the contractor (21% complete);
- Early warnings from the client (16% complete); and
- Project man-hours (37% complete).

Table 6-2 shows the percentage of missing values for each dataset. The last column of the table shows the percentage of missing information when the three features – that had the least amount of information available – were left out for the purposes of data mining.

Table 6-2: Dataset statistics with missing values

	Features missing values (%)	**Classification features missing values (%)**	**Features missing values after selected features left out**
Primary dataset	19.9%	20.8%	15.9%
Selected minutes	11.3%	0.5%	6.4%

The following trends were noted by the author when capturing the data from the minutes into the data warehouse:

- The project completion percentage was often not recorded during the progress meetings. In many cases, the individual structures or areas were mentioned along with progress on those structures, but the percentages of the structures or complete project

were not included. From a third-party view, this makes tracking progress over time or judging how well a project was doing very difficult. The earned-value percentage of the project and the planned completion percentage should be recorded. Both of these values are obtainable when using planning software such as MS Projects, Primavera or Candy.

- A few of the meeting minutes indicated that the programme had not been submitted by the second progress meeting (or in some cases submitted and not approved) which could indicate a lack of planning and/or commitment from both sides.
- A common theme was that the expected completion dates shown on minutes changed with every progress meeting, more so when the expected completion date was later than the contractual completion date.
- The actual contractual completion date was not always clearly available.
- Payments from the client to the contractor (and, in some cases, payment of consultants) played a role in the progress and completion of certain projects. This was not taken into account as a separate feature for this study due to the confidentiality of the financial information as indicated in Section 1.3.

6.4. Data Mining

For the data mining step, two different data mining applications were used – Orange and RapidMiner. Both are data mining applications.

Orange is an open-source toolkit for data mining and data visualisation. It uses widgets in a visual-programming interface to allow the user to set tasks.

RapidMiner is a subscription-based software package designed for data analytics. It also uses a visual interface to create a workflow similar to that of Orange. One of the main features of the program is that it has an auto model function that uses machine learning to create a model based on the type of output selected. The auto model process is a guided process that simplifies the data preparation, model selection and feature engineering. The models that are created are annotated and documented to allow the user to understand what the application did. Each step of the process can also be tuned and refined as required.

The data mining process was repeated several times with the dataset to ensure stability of the results. The way in which the dataset was subdivided for testing and scoring was changed between each iteration. Two different sampling methods were used: cross validation with 20

folds, and random sampling (with 100 iterations randomly splitting the dataset between 66% training and 34% testing). The process was repeated for each of the three datasets identified in Section 4.4, excluding the text-only dataset (Dataset T). The dataset containing text was data mined for only sentiment analysis and word counts.

The following sections describe the creation of the prediction models in both Orange (Section 6.4.1) and RapidMiner (Section 6.4.2).

6.4.1. Choosing an Appropriate Model and Learner

As mentioned in Section 3.2, several different types of data mining methods and learners can be used. The difference between a model and a learner is that a model is the complete task or workflow, while a learner is the specific algorithm being used. A prediction model is created in the specific application through data mining historic information and by using that information to make predictions. An example of a prediction model is shown in Figure 6-3. The entire workflow is the prediction model as all the parts are required to create the prediction model. The learner is a part of the model and is coloured pink. The learner is the specific data mining algorithm used to "learn" from the historic data.

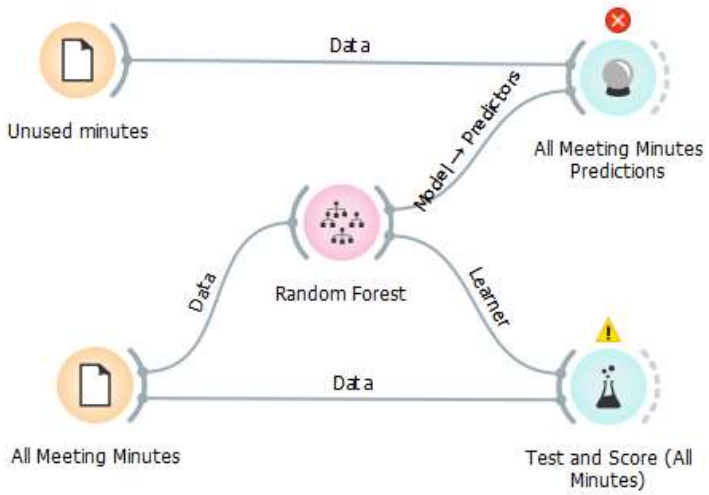


Figure 6-3: Prediction model in Orange

The type of model to be used depends on the information being mined. Because the information in Datasets A, F and S contained text as well as numbers, more than one type of learner was required so that all stored data could be used. The different modelling techniques used for this study, and a brief explanation for each, are discussed in the following sections.

Because this is the first time that site meeting minutes were used as input in a data mining process, all the data mining methods available in Orange and RapidMiner were used. As expected, some of the learners were not suitable for the dataset. Section 6.4.2 discusses the algorithms used to create the prediction models in Orange, while Section 6.4.3 discusses the algorithms used to create the prediction models in RapidMiner.

6.4.2. Orange Data Mining and Models

Various learners were applied to the datasets to determine which would yield the best results and which could provide new knowledge. The learners that were used in Orange are k-Nearest Neighbour (kNN), Logistic Regression, Linear Regression, Decision Trees, Random Forest, Neural Networks, Support Vector Machine and Naïve Bayes. The descriptions and uses of the learners were all obtained from the documentation on the Orange website (University of Ljubljana, 2020). Each of the different learners that were used to create models in Orange are discussed below, including an explanation of the learner and an example where possible.

Orange uses widgets to create the workflows. The user connects various widgets to create the workflow. Figure 6-4 shows an example of the workflow created for testing the selected meeting minutes dataset. The pink-coloured widgets that are directly linked to the file widget are all the learners (algorithms) that were used to create different models. In this case, seven models were created simultaneously. This functionality allows the user to compare the models directly with one another.

All the widgets are connected in sequence. The prediction model(s) start from the data file, which is connected to the learner widgets and, subsequently to a test-and-score widget. The test-and-score widget “tests” the different models that were created by the learners. This is done by sampling the data and providing feedback to the learner about its accuracy. This process is then repeated several times to improve the accuracy of the learner. Finally, the learners and the test-and-score widget are connected to the prediction widget. This completes the model which can then be used to make predictions on new datasets. An example of a complete prediction model is shown in Figure 6-3. The model uses only one learner (algorithm) to create the prediction model.

All the features used for data mining must be allocated a role and a type. Appendix 4 shows a screenshot of the file widget with the roles, types and feature names used for data mining. It

is important to assign a target feature which will be used to create the models. For this study, the target feature was whether the project finished late (yes/no).

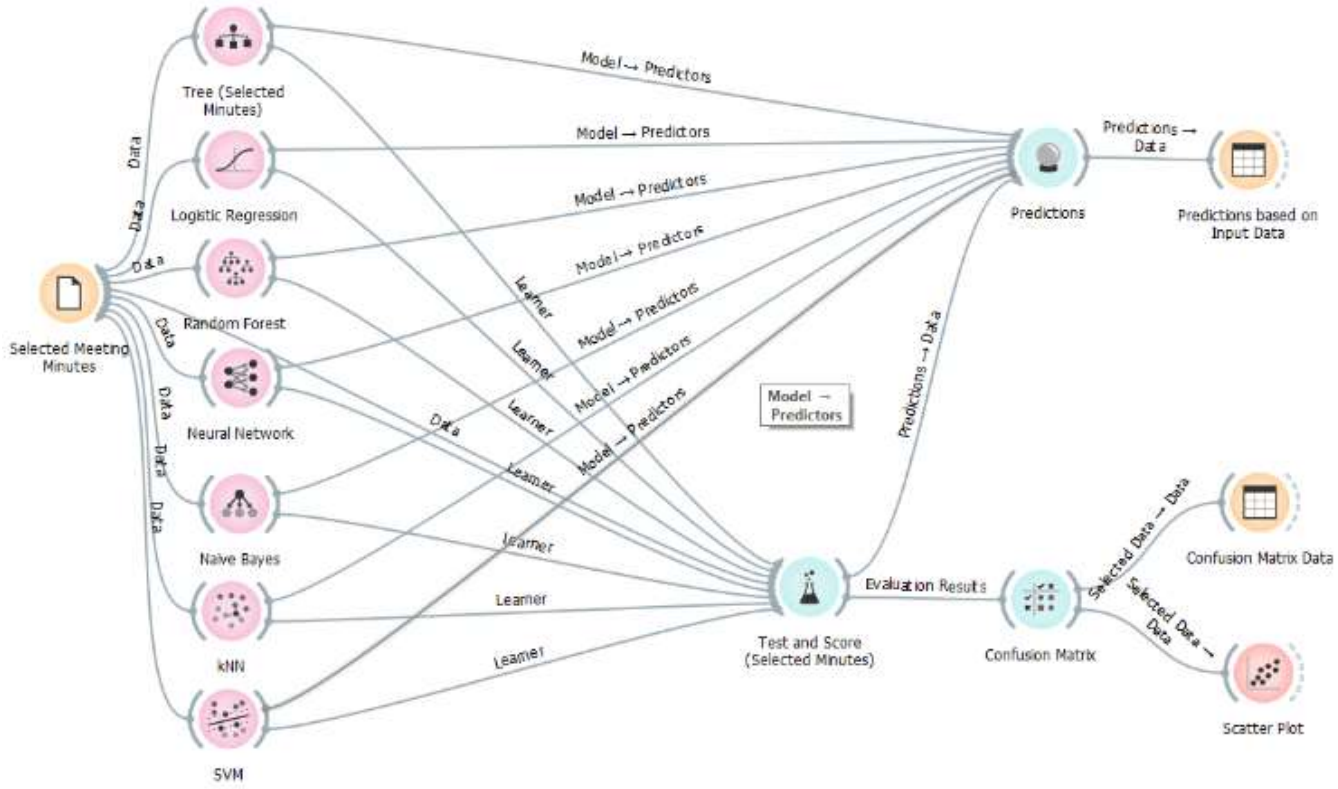


Figure 6-4: Example of Orange workflow created for testing of selected meeting minutes

6.4.2.1. *Orange Prediction Algorithms*

kNN: The kNN, or k-Nearest Neighbour, learner is a non-parametric method and uses the kNN algorithm to search for k-closest training examples in feature space and uses their average to do the predictions. The metric and the weights can be set within the widget. The metric can be Euclidean, Manhattan, Maximal or Mahalanobis. The weights can be set as uniform or distance. The closer the neighbours are, the greater the influence they have. The letter "k" in kNN stands for the number of nearest neighbours used in the classification. For this study, the k-value was set to the default which is calculated by the square root of the number of instances.

An example of how the learner works is shown below. In the example, there are four rows of information for a product in the dataset (Type 1 to 4), each with three features (Table 6-3). The kNN learner is used to determine what the probability is that the unknown product turns out to be good.

The distance between the unknown data and the known data for each row is calculated. If the Euclidian metric is used, the distance is calculated from the Euclidean equation:

$$d(p,q) = d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

where,

q_i = criterion 1 for case i;

p_i = criterion 2 for case i.

The dataset is then ranked according to the distances that were calculated. The distance of the unknown data is calculated and compared to the known information to determine its closest neighbour.

In this example, the k-value was chosen to be 3. If k=3, the probability that the unknown product is good is high, because there are two good products and one bad product in the ranking (an average of 66.67%). When comparing the information in Table 6-3 graphically, the three closest neighbours to the unknown orange dot are easily spotted (circled in red) in Figure 6-5.

Table 6-3: kNN Learner example

Information provided in dataset				**Information calculated**	
Name	**Criterion 1**	**Criterion 2**	**Class**	**Distance**	**Rank**
Type 1	7	7	Bad	4	3
Type 2	7	4	Bad	5	4
Type 3	3	4	Good	3	1
Type 4	1	4	Good	3.6	2
Unknown	3	7			

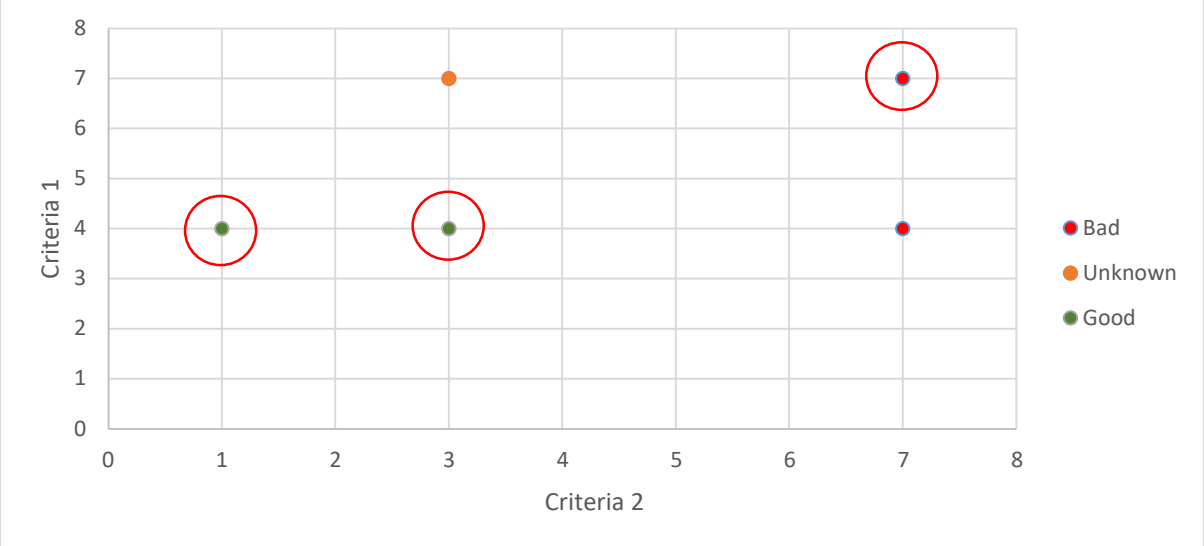


Figure 6-5: kNN example graph

Logistic Regression: The learner uses a logistic regression algorithm with LASSO (L1) or ridge (L2) regularisation to classify data. It can be used for classification of tasks only and is appropriate to use if the dependant variable is binary. LASSO regression is used to minimise a penalised version of the least squared function with the L1-norm penalty, while the ridge regularisation uses the L2-norm penalty.

The equation for logistic regression can be written as:

$$P(Y_i) = \frac{1}{1 + e^{-(b_0 + b_1 X_i)}}$$

where

- $P(Y_i)$ is the predicted probability that Y is true for case i;
- b_0 is a constant estimated from the data;
- b_1 is a b-coefficient estimated from the date;
- X_i is the observed score on variable X for case i.

Logistic regression is about estimating the values for b_0 and b_1 through maximum likelihood estimations. The values cannot be readily computed and must be estimated through multiple iterations on a computer.

The example data for the logistic regression learner is shown in Figure 6-6. The graph shows the results from a shooting test. A single shot was fired at each distance. It was recorded whether the target was hit (marked as 1) or missed (marked as 0) with blue squares. The

orange circles are the logistic-regression-calculation values for the probability of hitting the target at the various distances.

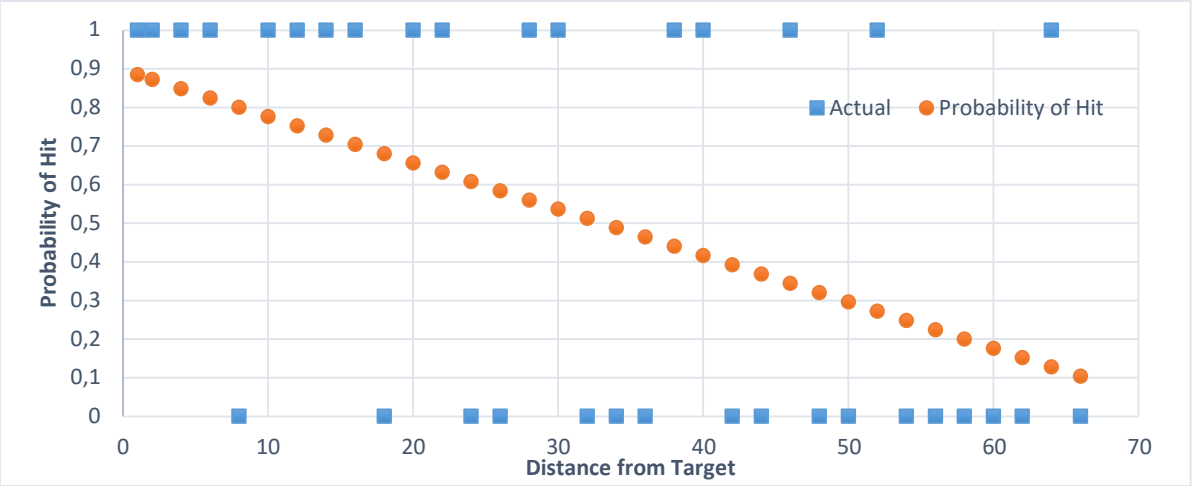


Figure 6-6: Example scatter graph for logistic regression

Linear Regression: The widget uses a linear regression model to construct a learner or predictor, that learns a linear function with optional LASSO (L1), ridge (L2) or elastic-net (L1L2) regularisation. The model identifies the direct relationship between the predictor variable and the response variable. The elastic-net regularisation (L1L2) is a mix of the other two types of regularisation and the ratio of the mix can be adjusted in the widget interface. Since the data used did not fit a linear model, this model could not be created.

The equation for linear regression is the same as the equation for a straight line:

$$Y = a + bX$$

An example of linear regression is shown in Figure 6-7, where a straight line was fitted in a dataset.

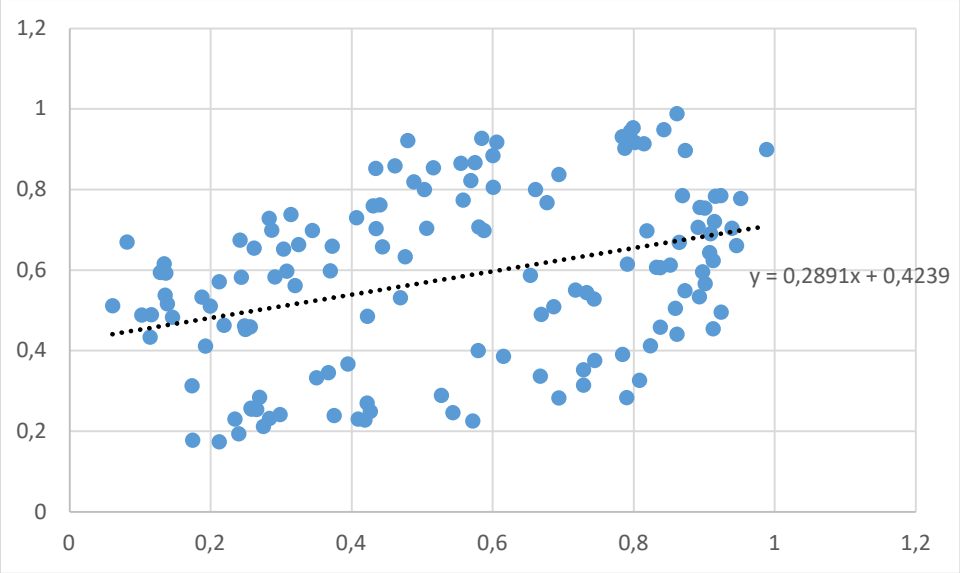


Figure 6-7: Linear regression example

Tree: The tree algorithm (more commonly known as decision trees) is a simple algorithm that splits the data into various nodes and can be used for both classification and regression tasks. The decision tree can handle multidimensional data and does not require domain knowledge. The tree is a pre-cursor to the random forest algorithm. Figure 6-8 and Figure 6-9 show the decision trees created by Orange for Datasets S and A, respectively.

The tree algorithm(s) decide(s) which feature is the most significant attribute in the dataset, which it then uses to split the dataset into subsets. The selected attribute is designated as the root node. The nodes are split on all the available variables, and the algorithm selects the split which results in the most homogeneous sub-nodes. In other words, the information gain is calculated for each node, and the nodes with the highest gains are chosen. If the tree that was created has many nodes, it might represent noisy data or outliers. Unwanted data can be removed by pruning the tree, which can increase the accuracy of the prediction.

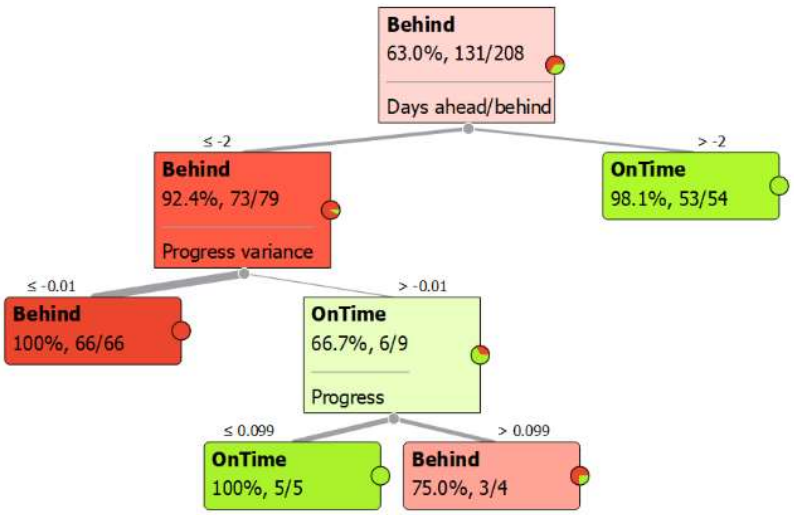


Figure 6-8: Decision tree created for Dataset S

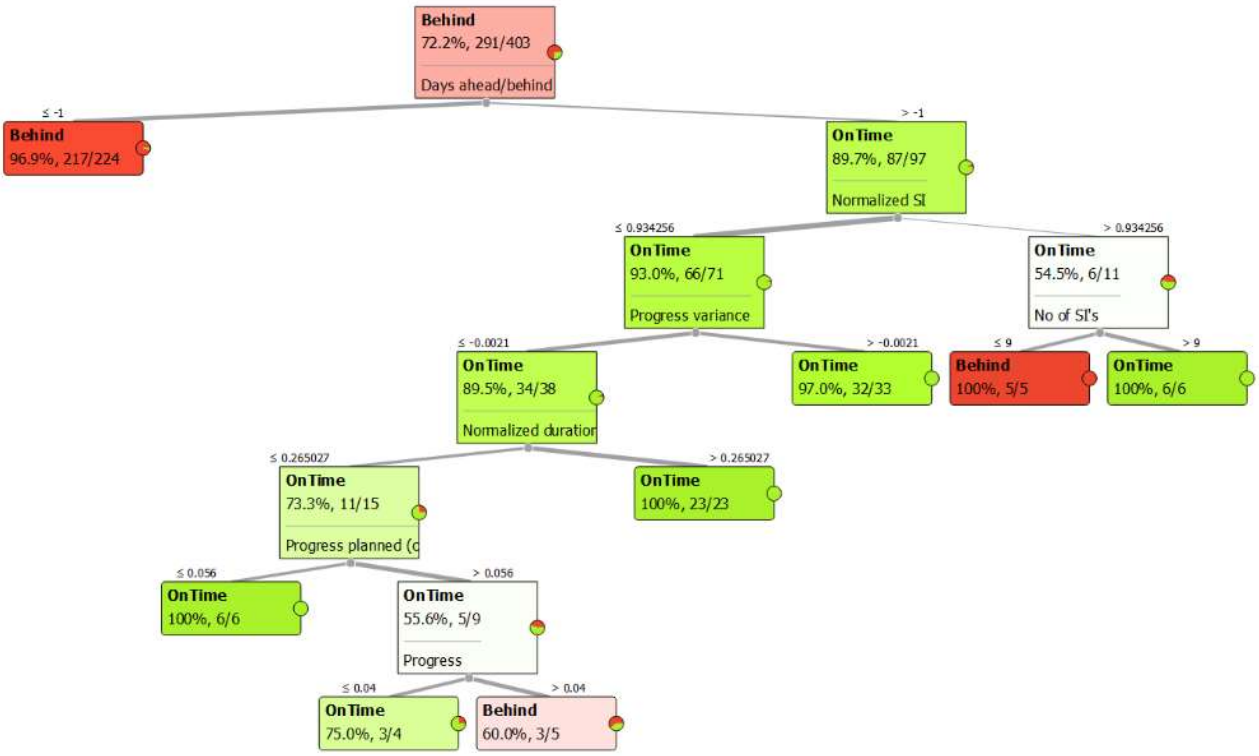


Figure 6-9: Decision tree created for Dataset A

Random Forest: A random forest is a group of decision trees. Each tree is generated from a data sample. This is then used to select the best feature for the split. The final model is based on the majority vote from the individually developed trees in the forest. The algorithm determines how many of the trees provide the same (or similar) answer and, thus, the majority "vote" creates the final model. Random forests can be used for both classification and regression tasks.

Figure 6-10 shows 10 decision trees created in Orange that form the random forest for Dataset A. The number of trees created is not fixed and depends on the data and algorithm which, in this case, produced 10 trees for the prediction model using Dataset A.

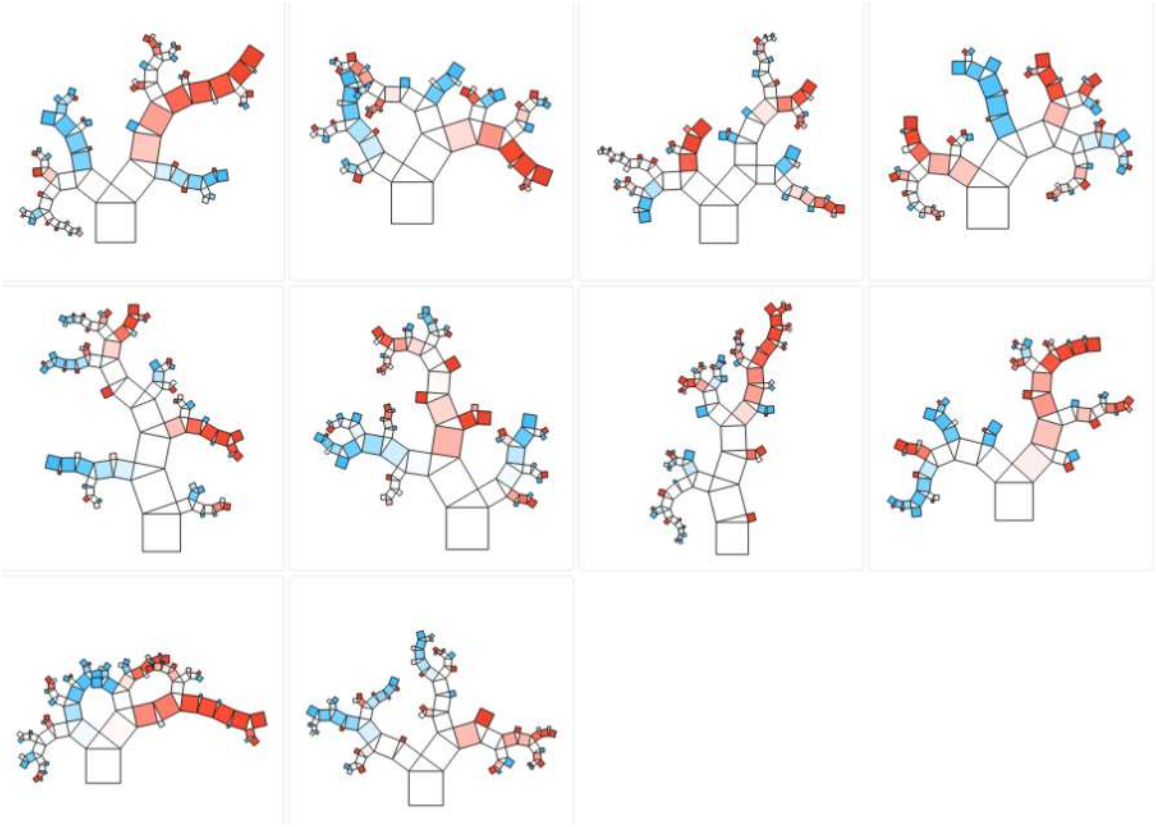


Figure 6-10: Random forest created for Dataset A in Orange

Neural Network: A neural network algorithm is used to create models that are linear or non-linear. Neural networks recognise hidden patterns in raw data by creating interconnected nodes, much like neurons in human brains. Information is fed into the network via the input neurons which, in turn, triggers the layers of hidden neurons. The information then arrives at the output neurons. Feedback is required as this is how the neural network learns what it has done right and wrong. This is normally done through a process called back-propagation, where the output of the network is measured and compared to the output it was expected to deliver. The input is accompanied by the label explaining what should have been delivered to the output. Once the difference is measured, the weights of the connections between the neurons are adjusted backwards – from the output to the hidden neurons and, finally, to the input neurons. Through time, the network learns by narrowing the difference in outputs until

the two outputs match exactly. Figure 6-11 shows a neural network with two layers of hidden neurons.

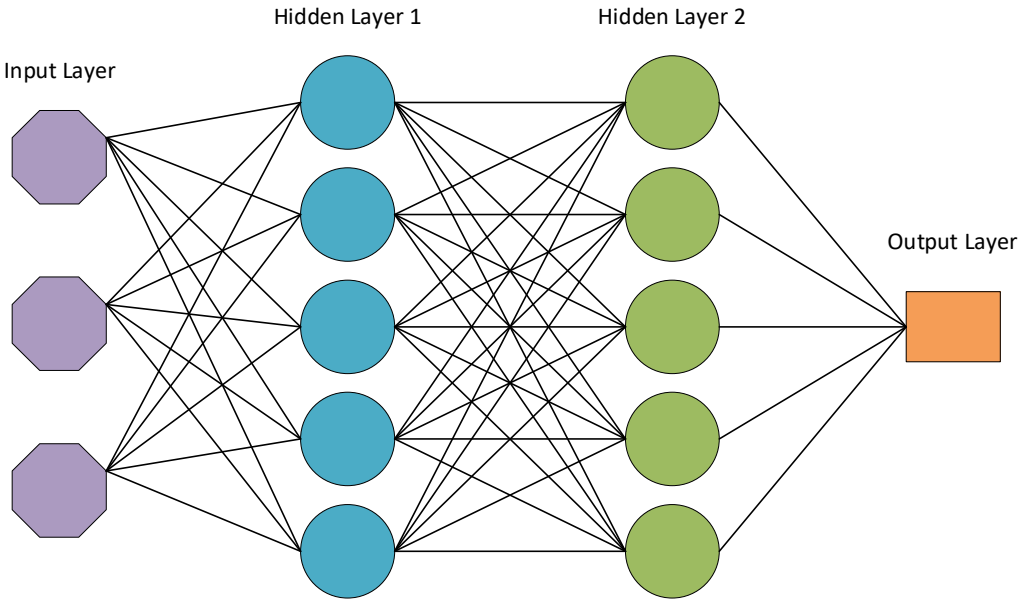


Figure 6-11: Neural network example

Support Vector Machine: The Support Vector Machine (SVM) learner is a classification and regression algorithm based on the LibSVM package. The Linear SVM learners are more suitable for large-scale problems because they are fast.

SVM fits a line between two sets of data points to maximise the space between them. This is done with a hyper plane. If the data is two-dimensional in nature, the hyper plane will be created in a third dimension where the plane is flat, and a line can be drawn between the sets of data.

A dataset, as shown in Figure 6-12, was created to explain this. A straight line cannot be drawn between the two sets of data to maximise the space between them. The data must be transformed to a higher dimension. If a third dimension is added, where $x^2+y^2=z$, the dataset can be transformed into Figure 6-13. Value z represents the cube of the distance from the origin. A straight line can be drawn between the two datasets as shown by the red dotted line in Figure 6-13. Since $x^2+y^2=z$ is the equation of a circle, the linear line can be projected onto the dataset as a circle as shown by the red dotted line in Figure 6-12.

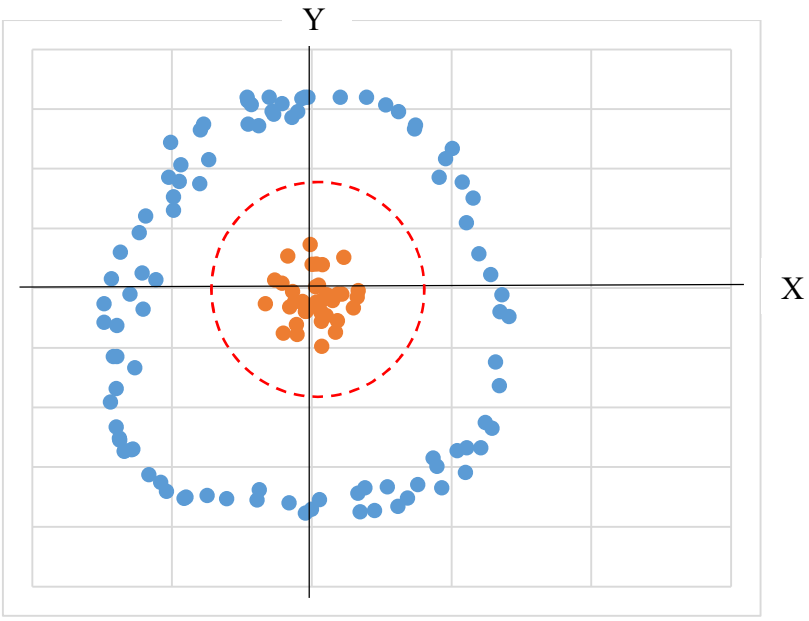


Figure 6-12: Dataset created for SVM example

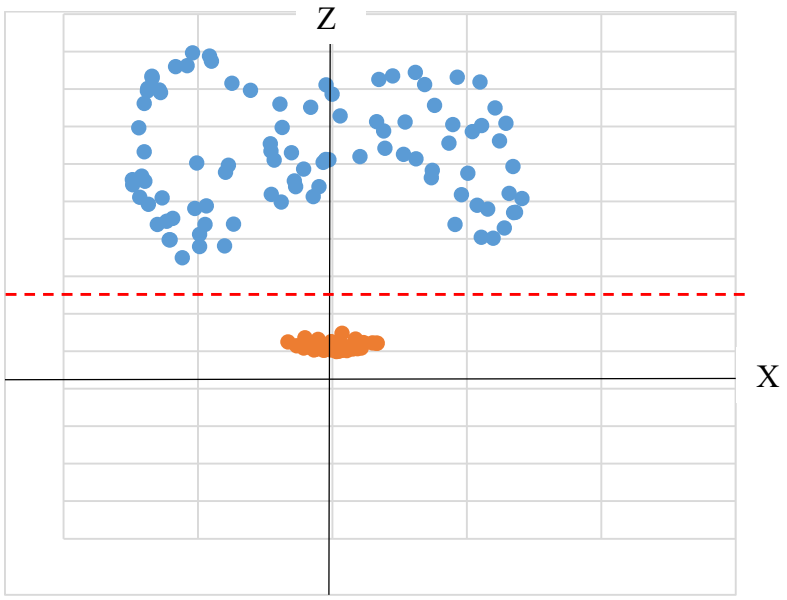


Figure 6-13: SVM higher-dimension data

Naïve Bayes: The algorithm uses the Naïve Bayesian model to learn from the data. It is used for probabilistic classification tasks with the assumption of feature independence.

The algorithm uses Bayes' theorem to calculate the conditional probability. Naïve refers to the fact that it assumes that all the features are independent even if they are not. The Bayes Naïve probability equation can be written as:

$$Probability\ of\ outcome = \frac{Prior\ probability * \ Probability\ of\ likelihood}{Probability\ of\ evidence}$$

The following example was adapted from the example of Prabhakaran (2018). The prediction algorithm should predict the probability that an animal is a mammal. The same method is used to determine the probability that an animal is not a mammal, and the two results are compared. The highest result is the most probable. The dataset is shown in Table 6-4.

Table 6-4: Dataset for Naive Bayes example

Type	Predator	Not predator	Toothed	Not toothed	Tail	No tail	Total (Each Pair)
Mammal	300	200	375	125	400	100	500
Bird	0	300	150	150	300	0	300
Fish	100	100	125	75	125	75	200
Total	400	600	650	350	825	175	1000

From the dataset, the prior probability that an animal is a mammal is calculated as

$$\text{Probability}(B=\text{Mammal}) = 500/1000 = 0.50$$

The probability of evidence is the same for all the classes, and is calculated as

$$P(A_1=\text{Predator}) = 400/1000 = 0.40$$

$$P(A_2=\text{Toothed}) = 650/1000 = 0.65$$

$$P(A_3=\text{Tail}) = 825/1000 = 0.825$$

The probability of the animal being a mammal is calculated as

$$P(A_1=\text{Predator} \mid B=\text{Mammal}) = 300/500 = 0.60$$

$$P(A_2=\text{Toothed} \mid B=\text{Mammal}) = 375/500 = 0.75$$

$$P(A_3=\text{Tail} \mid B=\text{Mammal}) = 400/500 = 0.80$$

The overall probability of evidence for the animal being a mammal = 0.6 * 0.75 * 0.8 = 0.36.

Substituting the values into the Bayes equation provides:

$$P(Mammal \mid \text{Predator, Toothed and Tail})$$
$$= \frac{P(Predator \mid \text{Mammal}) * P(Toothed \mid \text{Mammal}) * P(Tail \mid \text{Mammal}) * P(Mammal)}{P(Predator) * P(Toothed) * P(Tail)}$$
$$= \frac{0.6 * 0.75 * 0.85 * 0.5}{0.4 * 0.65 * 0.825} = 0.839 \; or \; 83.9\%$$

6.4.3. RapidMiner Data Mining and Models

The different learners that were used to create models in RapidMiner were Deep Learning, Fast Large Margin, Generalised Linear Model, Gradient Boosted Trees, Logistic Regression, Naïve Bayes, Random Forest and SVM. The description of the models used in RapidMiner was obtained from the documentation on their website (RapidMiner, 2020).

RapidMiner uses operators to create the workflow. Figure 6-14 shows a typical workflow created in RapidMiner for this study. The figure shows one of the workflows created by the Auto Model feature. There are several steps for creating a prediction model in RapidMiner. Each of these steps has workflows that manage specific parts of the data mining process. The steps for creating the prediction model are: (1) pre-processing; (2) feature engineering; (3) transform validation and scoring data; (4) scoring, validation, explanations, weights and simulator; (5) production model; and (6) process results.

Each of the different learners that were used to create models in RapidMiner is discussed below. Most of the algorithms are similar to Orange, therefore, no examples are given.

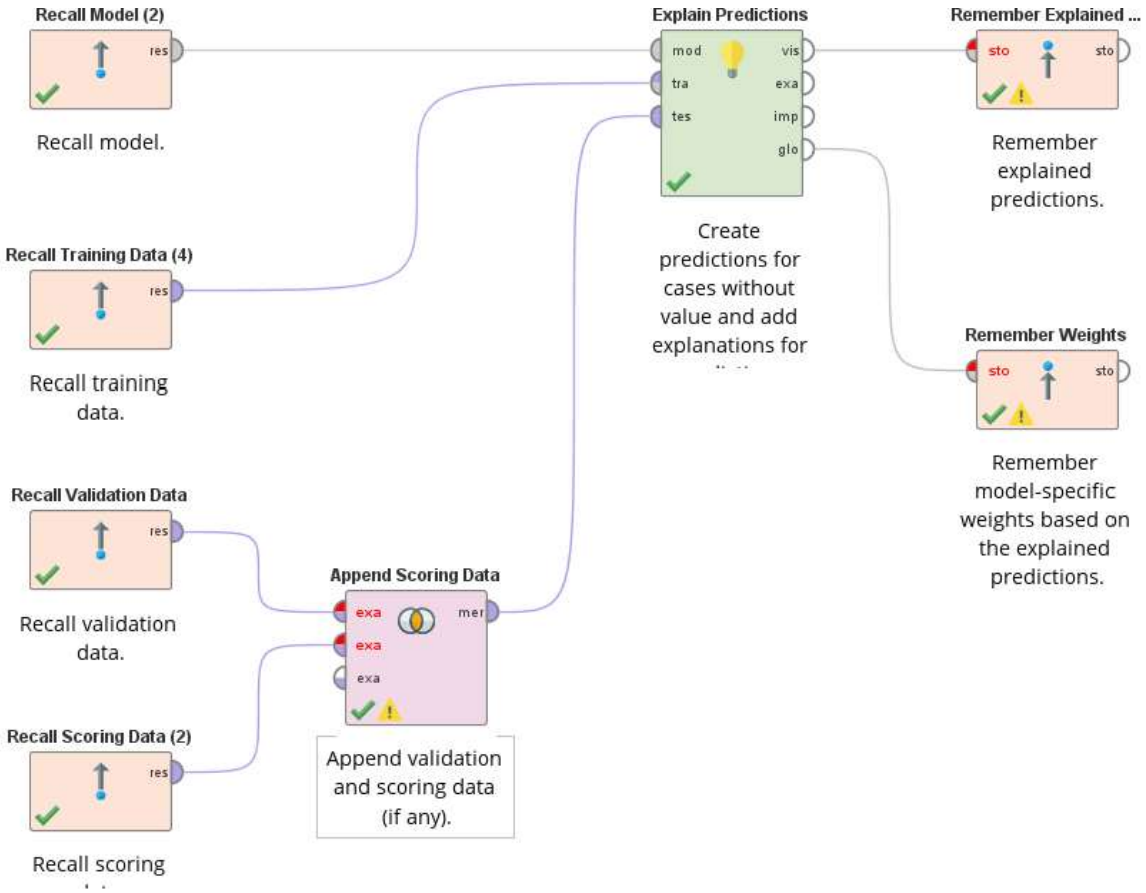


Figure 6-14: RapidMiner workflow

6.4.3.1. *RapidMiner Prediction Algorithms*

Deep Learning: The deep learning algorithm in RapidMiner is similar to the neural networks algorithm, but taken to another level. It is based on multi-layer feed-forward artificial networks. It often needs huge amounts of training data and has a large number of parameters to tune. It can, however, be used on smaller datasets if needed. Running it can require a big investment in computing power, as it is often not flexible and is domain-specific.

Fast Large Margin: The fast large margin algorithm applies a SVM algorithm. It takes the input data and predicts in which of the two possible categories each given input into will fit. It is, thus, a binary, non-probabilistic classifier.

Generalised Linear Model: This is an extension of the traditional linear-model algorithm where the algorithm fits generalized linear models to the data by maximising the log-likelihood. This is similar to the linear regression algorithm in Orange where the elastic-net penalty can be used for parameter regularisation.

Gradient Boosted Trees: This algorithm is a forward-learning ensemble of either regression or classification trees It is used to obtain predictive results through gradually improved estimations. Figure 6-15 is the decision tree generated by the learner during the mining process.

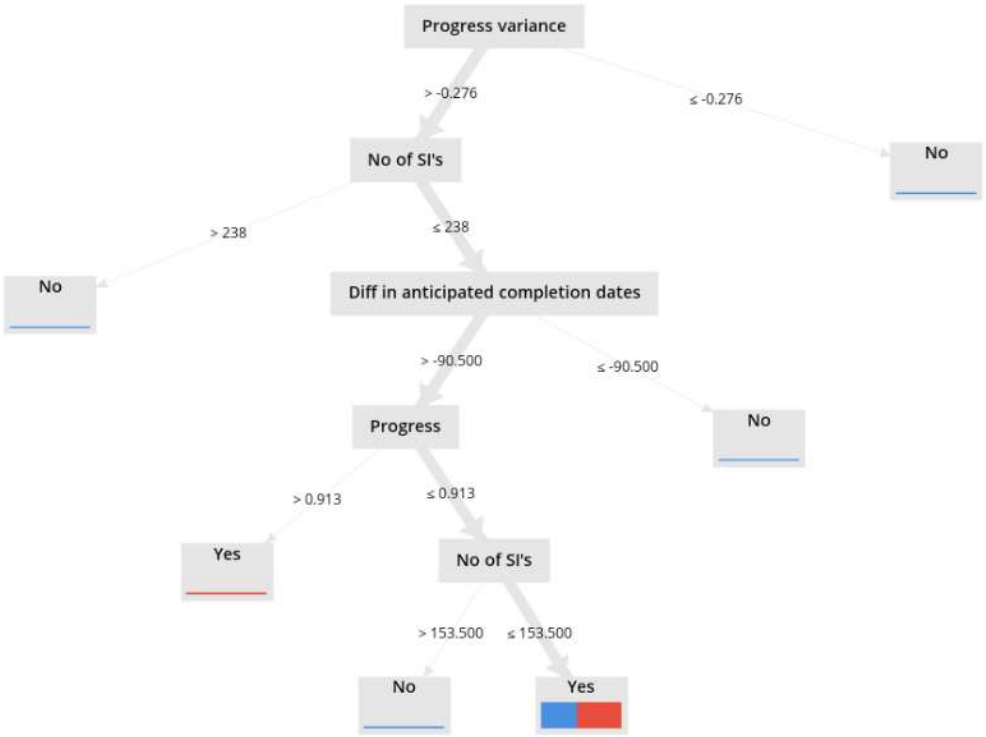


Figure 6-15: Decision tree from RapidMiner learner

Logistic Regression: This learning method can be used for classification and regression. It is based on the Java implementation of *myKLR* by Stefan Rueping (RapidMiner, 2020).

Naïve Bayes: This algorithm works similar to the one in Orange and is used for building a classification model.

Random Forest: This algorithm is similar to the random forest algorithm in Orange. Figure 6-16 shows a decision tree generated during this process. Sixty trees were created from Dataset A.

Support Vector Machine: This learner is created by the Java implementation of *mySVM* by Stefan Rueping (RapidMiner, 2020). It can be used for both regression and classification. It is a fast algorithm and provides good results for many different learning tasks. It works with asymmetric, linear and quadratic functions. The SVM works as per the description in the Fast-Large Margin operator.

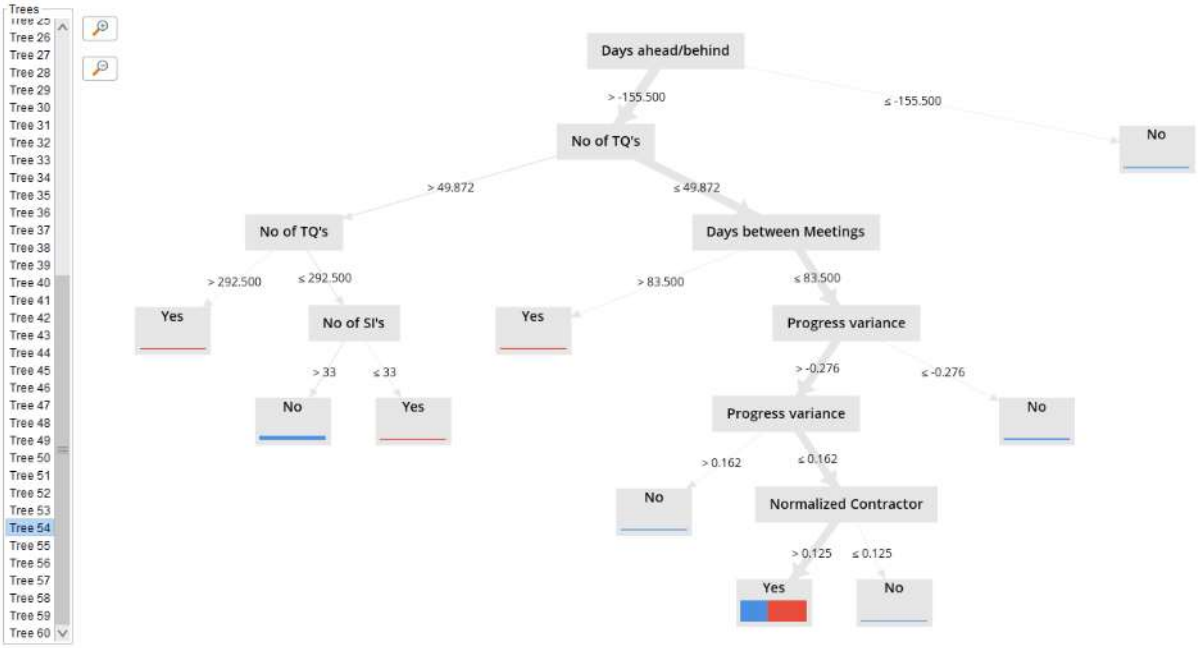


Figure 6-16: Decision tree generated by random forest learner in RapidMiner from Dataset A

6.5. Data Analysis

The analysis of the predication models created is discussed in this section. The analysis of the prediction models created in Orange is discussed in Section 6.5.2, while the analysis of the prediction models created in RapidMiner is discussed in Section 6.5.3.

There are four common data mining topics discussed in Section 6.5.1. These topics are the same for both Orange and RapidMiner and do not need to be discussed separately. The common topics are the confusion matrix (Section 6.5.1.1), correlations (Section 6.5.1.2), text mining and sentiment analysis (Section 6.5.1.3) and model gains (Section 6.5.1.4).

6.5.1. Analysis of Common Data Mining Topics

6.5.1.1. *Confusion Matrix*

The confusion matrix is created to provide the data analyst with feedback on how the application predicts the outcome of the dependant variable. It provides the analyst with a matrix of actual values compared to the predicted values when verifying the model. Since the confusion matrix does the same work in both Orange and RapidMiner, it is discussed under the common data mining topics section.

The output of a typical confusion matrix created for Dataset S is shown in Figure 6-17 with a typical data table of misclassified features shown in Figure 6-18.

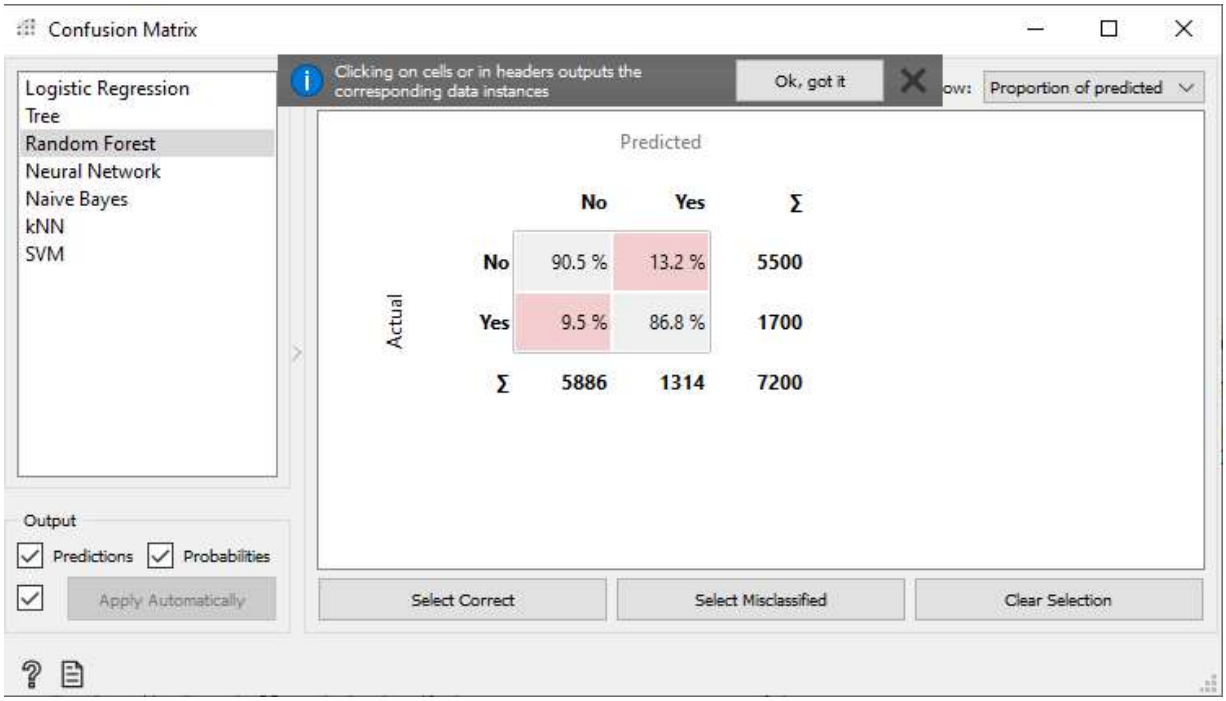


Figure 6-17: Confusion matrix created for the random forest prediction model for Dataset S in Orange

In Figure 6-17, the confusion matrix shows the number of actual project outcomes compared to the number of predicted project outcomes. Using the results from testing the model, 7200 predictions were made and then compared to the actual outcomes recorded for each prediction (34% of Dataset S repeated 100 times). The random forest model was able to predict when the project would be completed behind schedule 86.8% of the time and when it would be completed ahead of time or on schedule 90.5% of the time. It is, thus, slightly more accurate in predicting project completion ahead or on time than it is in predicting the late completion of a project. It was wrong in 9.5% of the predictions where the model indicated that the project would not finish behind schedule but, in reality, it did. The model was wrong in its predictions 13.2% of the time where it indicated that the project would finish behind schedule but, in reality, it did not.

This same trend continues for all the other models created. There are two exceptions: the Naïve Bays model, that is slightly more accurate in predicting whether a project will finish on time (90.8%); and the SVM model, that is more accurate in predicting whether a project will finish behind schedule (92.1%).

Upon examination of the confusion matrix created for the most accurate prediction model, the Random Forest, the following was seen: Orange bases its calculations of the probability for

on-time project completion on the premise that the progress percentage will be higher than the planned progress percentage. The meeting minutes, however, indicated that the project was behind schedule. Thus, it was not an accurate prediction. Several other inaccurate predictions indicated behind-schedule progress where the reverse was true. There were similar instances where a project was behind – based on the recorded progress percentage and the days actually recorded – but the meeting minutes recorded the project as being on time because the forecast for the project showed that the milestone dates would be met.

Figure 6-18 is an extract of the misclassified predictions from the random forest model. The first column is the actual outcome of that specific project based on the meeting minutes. The second column is the predicted outcome. Columns 3 and 4 are the probabilities for the outcome No (finishing on time) and Yes (finishing behind schedule). The remainder of the columns are some of the features/attributes recorded for that specific meeting's minutes. The coloured bars show the weights that the model placed on a specific feature contributing towards the prediction. Red indicates a No outcome and blue indicates a Yes outcome. The figure shows that, for the majority of the outcome predictions in this extract, the model relied heavily on the progress and progress planned features of the dataset.

sh Late? (Yes/No/Unsı	(Yes/No/Unsure)(Ran	p(No)	p(Yes)	Progress	Progress planned
Yes	No	1	0	0.89	0.88
Yes	No	0.6	0.4	0.61	0.79
Yes	No	0.766667	0.233333	0.87	1
Yes	No	0.6	0.4	0.72	0.89
Yes	No	0.8	0.2	0.86	0.99
Yes	No	0.59	0.41	0.11	0.11
Yes	No	0.783333	0.216667	0.69	0.84
No	Yes	0.375	0.625	0.97	0.98
Yes	No	0.8	0.2	0.0239	0.15
No	Yes	0.4	0.6	0.897	0.9545
No	Yes	0.380556	0.619444	1	1
No	Yes	0.48	0.52	0.0156	0.0278
No	Yes	0.44	0.56	0.08	0.08
Yes	No	0.811429	0.188571	0	0
Yes	No	0.5	0.5	0.61	0.79
No	Yes	0.441667	0.558333	0.01	0.01
Yes	No	0.708333	0.291667	0	0

Figure 6-18: Misclassified data table output created from confusion matrix for Dataset S in Orange

6.5.1.2. *Correlations*

The correlations in the data were checked for any obvious correlations between the different features. This was done in both Orange and RapidMiner but, since the correlation data is a feature of the data and not of the programs themselves, the correlations are only discussed in this section. The correlations calculated by Orange and RapidMiner are not exactly the same but are in a similar order of magnitude. For comparison, the correlations from Orange are shown in brackets after the correlations from RapidMiner in the paragraph below. Some of the obvious correlations included the correlation of the meeting number with the duration of the project. This indicates that the information was processed correctly. An extract of the correlations table from the data mining process is shown Figure 6-19 with the aforementioned correlation highlighted for information. RapidMiner automatically highlights a cell with a darker shade when the correlations for the pair of features are higher.

A correlation is always between -1 and +1. A correlation of **+1** means that there is a perfect positive and linear relationship, while a correlation of **-1** means a perfect negative and linear relationship. A correlation of **0** shows no relationship at all. A correlation of **0.5** means that there is a moderate relationship between the variables.

Some of the correlations noted from RapidMiner are:

- A correlation of 0.533 (0.450) between compensation events and site instructions – This shows that when more site instructions are issued to the contractor, there is a tendency for the contractor to submit more claims or compensation events. This may be because site instructions can change the scope of work or delay the contractor.
- A correlation of 0.495 (0.456) between programme revisions and compensation events – This could indicate that there is a link between claims being submitted and the programme being subsequently revised because of it.
- A correlation of 0.553 (0.564) between RFI's and project man-hours – This could be expected because the number of man-hours will increase as the duration of the project increases. The number of RFI's will also increase as the duration of the project increases, hence, the expectation of the correlation between the two features. This is confirmed with the correlations between the RFI's and site instructions, meeting numbers and duration of the project all being of the same order.

Attributes	C#CE	C#Days ...	C#Days ...	C#Diff i...	C#Norm...	C#Norm...	C#Norm...	C#Norm...	C#Norm...
C#CE	1	0.047	-0.146	-0.015	-0.088	0.032	0.405	0.398	0.533
C#Days ahead/behind	0.047	1	-0.251	-0.170	-0.107	0.025	-0.273	-0.270	-0.108
C#Days between Meetings	-0.146	-0.251	1	0.103	0.049	0.035	0.206	0.193	0.057
C#Diff in anticipated completion dates	-0.015	-0.170	0.103	1	-0.077	0.066	0.069	0.062	0.025
C#Normalized Client	-0.088	-0.107	0.049	-0.077	1	0.140	0.034	0.038	0.028
C#Normalized Contractor	0.032	0.025	0.035	0.066	0.140	1	-0.170	-0.170	-0.063
C#Normalized duration	0.405	-0.273	0.206	0.069	0.034	-0.170	1	0.987	0.742
C#Normalized meeting number	0.398	-0.270	0.193	0.062	0.038	-0.170	0.987	1	0.763
C#Normalized SI	0.533	-0.108	0.057	0.025	0.028	-0.063	0.742	0.763	1

Figure 6-19: Example of the correlations obtained from RapidMiner

6.5.1.3. *Text Mining and Sentiment Analysis*

The text in the site meeting minutes was isolated into a separate dataset (Dataset T). This was to ensure that only the text is analysed, not other information entered into the dataset. This "text only" dataset was data mined to determine the sentiment of the individual site meetings. Some interesting trends from this shows that, when the project is completed late, the general sentiment in the site meeting minutes is negative. The reverse is true if the project was completed on or ahead of time. This can be expected when a project is expected to complete late as everyone working on the project will know this.

The text was also data mined to determine if any new knowledge can be learned from it. A word cloud is generated to indicate the frequency of words appearing in the text. From this, the most-used terms are programme, progress and safety. This is normal for the site meetings, and no new knowledge could be learned from the word cloud and data mining of the text in general. The word cloud that was generated in Orange is shown in Figure 6-20.

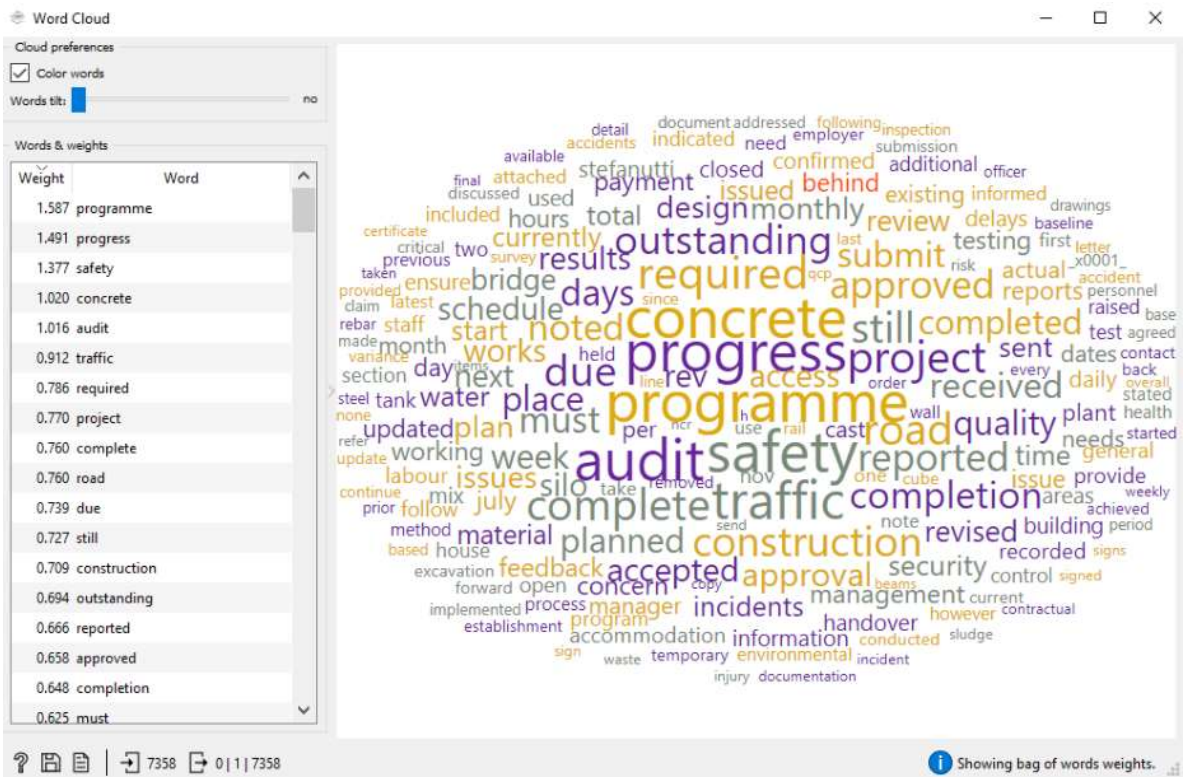


Figure 6-20: Word cloud generated in Orange for Dataset T

Figure 6-21 shows the typical output in the form of a table indicating the sentiment analysis of the dataset. The negative values show negative sentiment, while positive values show positive sentiment. There are exceptions to this. From an analysis of the exceptions, they seem likely to occur where the minutes are written in short sentences and bullet points.

The first column of the figure is the target feature (in other words, project completion on time or behind schedule). The next five columns are the five text features recorded from the site meetings. The last column shows the sentiment analysis of the meeting minutes. The red, blue and grey bars show whether project completion was late, on time or if no information was recorded. The length of the coloured bars is a visualisation of the numeric values.

	sh Late? (Yes/No/Unsu	Safety, Enviro & Com	Security/IR	Quality	ontractual/Commerci	Progress Notes	sentiment
1	Yes	Nothing to Report	Nothing to Report	Nothing to Report	Nothing to Report	Nothing to Report	0
2	Yes	Operator's licensing ...	Steelworkers strike is...	RZ enquired whethe...	Contractual issuesJ...	Stefanutti Socks to i...	-1.54525
3	Yes	Health & Safety Me...	Steelworkers strike is...	6.1 HM enquired ab...	Contractual issuesJ...	8.1 SS to issue updat...	-1.36674
4	Yes	Team from Shondan...	gplants causing dela...	6.1 SS to leave a tea...	7.1 HM suggested t...	8.1 SS to issue updat...	-0.2
5	Yes	Team from Shondan...	5.1 Steelworkers stri...	6.1 SS to leave a tea...	7.1 EM and AC meet...	8.1 SS to issue updat...	-0.325733
6	Yes	Last team arrived an...	5.1 Steelworkers stri...	6.1 SS to submit dat...	7.1 EM and AC meet...	8.1 SS to issue updat...	-0.808314
7	Yes	Last team arrived an...	5.1 Steelworkers stri...	6.1 SS to submit dat...	7.1 EM and AC meet...	8.1 SS to issue updat...	-0.538213
8	Yes	4.1 Health & Safety ...	Steelworkers strike o...	6.1 SS to submit dat...	7.1 EM and AC meet...	8.1 SS to issue updat...	-0.106952
9	Yes	4.1 Health & Safety ...	5.1 Nothing to record	6.1 An updated cub...	7.1 EM and AC meet...	8.1 SS to issue updat...	0.100604
10	Yes	4.1 Health & Safety ...	5.1 Nothing to record	6.1 An updated cub...	7.1 EM and AC meet...	8.1 SS to issue updat...	-0.369344
11	Yes	Nothing to report	Nothing to report	Nothing to report	Nothing to report	Nothing to report	0
12	Yes	4.1 Health & Safety ...	5.1 Nothing to record	6.1 An updated cub...	7.1 EM and AC meet...	8.1 SS to issue updat...	-0.493827
13	Yes	4.1 Health & Safety ...	5.1 Nothing to record	6.1 An updated cub...	7.1 EM and AC meet...	8.1 SS to issue updat...	-0.649351
14	Yes	4.1 Health & Safety ...	Nothing to record	6.1 An updated cub...	7.1 EM and AC meet...	8.1 SS to issue updat...	-0.479781
15	Yes	4.1 Health & Safety ...	Nothing to record	6.1 An updated cub...	7.1 EM and AC meet...	8.1 SS to issue updat...	-0.217391
16	Yes	It will be acceptable ...	The Site Security,incl...	?	CONTRACTUAL DAT...	Nothing to report	0.294985
17	Yes	2.8 Accommodation...	Site Security2.22 The...	SPECIFICATIONS AN...	Nothing to report	Rev 01 Programme s...	0.170358
18	Yes	SPECIFICATIONS AN...	3.27 A second CLO ...	Separate quality me...	Nothing to report	Project on average 1...	0.232558
19	Yes	ENVIRONMENTAL ...	Site Security4.15 The...	TANDARD OF WORK...	Nothing to report	5 days, 32 days and ...	0.619835
20	Yes	House keeping5.22 ...	The Contractor has ...	No discussion	Nothing to report	Nothing to report	1.47493
21	Yes	House keeping6.20 ...	The Contractor has ...	No discussion	Nothing to report	Nothing to report	-1.29032
22	Yes	House keeping7.24 ...	The Contractor has ...	Work is done accord...	Nothing to report	Nothing to report	0
23	Yes	House keeping8.17 ...	Theft of Temporary ...	No discussion	Nothing to report	Nothing to report	-1.19048
24	Yes	House keeping9.18 ...	Theft of Temporary ...	The Engineer confir...	Nothing to report	Nothing to report	1.25
25	Yes	House keeping10.18...	heft of Temporary Tr...	The Engineer rejecte...	Nothing to report	Nothing to report	-1.47059
26	Yes	The Engineer rejecte...	The Engineer rejecte...	No discussion	Nothing to report	Nothing to report	-4.34783
27	Yes	House keeping12.16...	Theft of reinforcing ...	Concrete Works at R...	Nothing to report	Nothing to report	2.15054
28	Yes	House keeping13.17...	The AoT Warning Sy...	No discussion	Nothing to report	Nothing to report	2.08333
29	Yes	House keeping14.17...	The AoT Warning Sy...	The defective buildi...	Nothing to report	Nothing to report	-4
30	Yes	ouse keeping15.17 ...	The AoT Warning Sy...	The finishing of the ...	Nothing to report	Nothing to report	0.793651
31	Yes	House keeping16.19...	During the past mo...	The finishing of the ...	Nothing to report	Nothing to report	0.318471
32	Yes	Theft on site is an o...	House keeping17.18...	Edwin Kruger from S...	Nothing to report	Nothing to report	1.0989
33	Yes	House keeping18.17...	Theft on site is an o...	Edwin Kruger from S...	Nothing to report	Nothing to report	1.36054
34	Yes	No Discussion	No discussion	Concrete Handrails ...	Nothing to report	Nothing to report	0
35	No	Existing Medicals us...	Nothing to report	9.1 The SS quality co...	Contract signing – ...	Contractor to submi...	1.63132
36	No	3.1 Existing medical...	Nothing to report	8.1 The SS quality co...	5.1 The Contract Do...	Nothing to report	1.67401
37	No	3.1 The use of safety...	Nothing to report	8.1 Site concrete mi...	5.1 Start date 01/06/...	Nothing to report	0.666667
38	No	The use of safety har...	Nothing to report	Site concrete mix de...	Start date 01/06/201...	SS submitted a revis...	0.401606
39	No	3.1 The use of safety...	Nothing to report	8.1 Site concrete mi...	5.1 Start date 01/06/...	4 days late	1.46341
40	No	3.1 There have been ...	Nothing to report	8.1 The Mahikeng su...	5.1 Start date 01/06/...	6 days late	0.481928
41	No	3.1 There have been ...	Nothing to report	8.1 Afrisam have ide...	5.1 Start date 01/06/...	6 days wind delay	0.872093
42	No	3.1 There have been ...	Nothing to report	8.1 The quality of th...	5.1 Start date 01/06/...	6.5 days actual com...	0
43	?	• The co...	Nothing to report	Nothing to report	a. Stefan...	5.1.1 WE...	-0.124844
44	?	1. This is ...	Nothing to report	S has submitted the ...	1. PJ me...	1. SS sub...	0.930233
45	?	1. Radiati...	Nothing to report	1. Qualit...	1. Contra...	1. SS will ...	1.22449
46	?	1. SS indi...	Nothing to report	1. SS foll...	1. SS-RE ...	95% complete site cl...	0.33557

Figure 6-21: Example of sentiment analysis

6.5.1.4. *Model Gains*

Another important result to consider is the gains of each model. The gain of the model is a measure of how much better the model can be expected to perform than when doing predictions without the model. The gain of a model goes hand-in-hand with the lift chart. The gains achieved for each of the models in RapidMiner are shown in Table 6-5 while an example of a lift chart created in Orange is shown in Figure 6-22. The purple line in the

figure shows the gain for the random forest prediction model. The vertical axis shows the sensitivity of the prediction model while the horizontal axis shows the value 1- specificity.

Table 6-5: Gains achieved my models in RapidMiner

Model type	**Dataset A gains**
Gradient Boosted Trees	78
Random Forest	70
Deep Learning	64
Support Vector Machine	50
Decision Tree	46
Logistic Regression	46
Generalised Linear Model	44
Naïve Bayes	38
Fast Large Margin	28

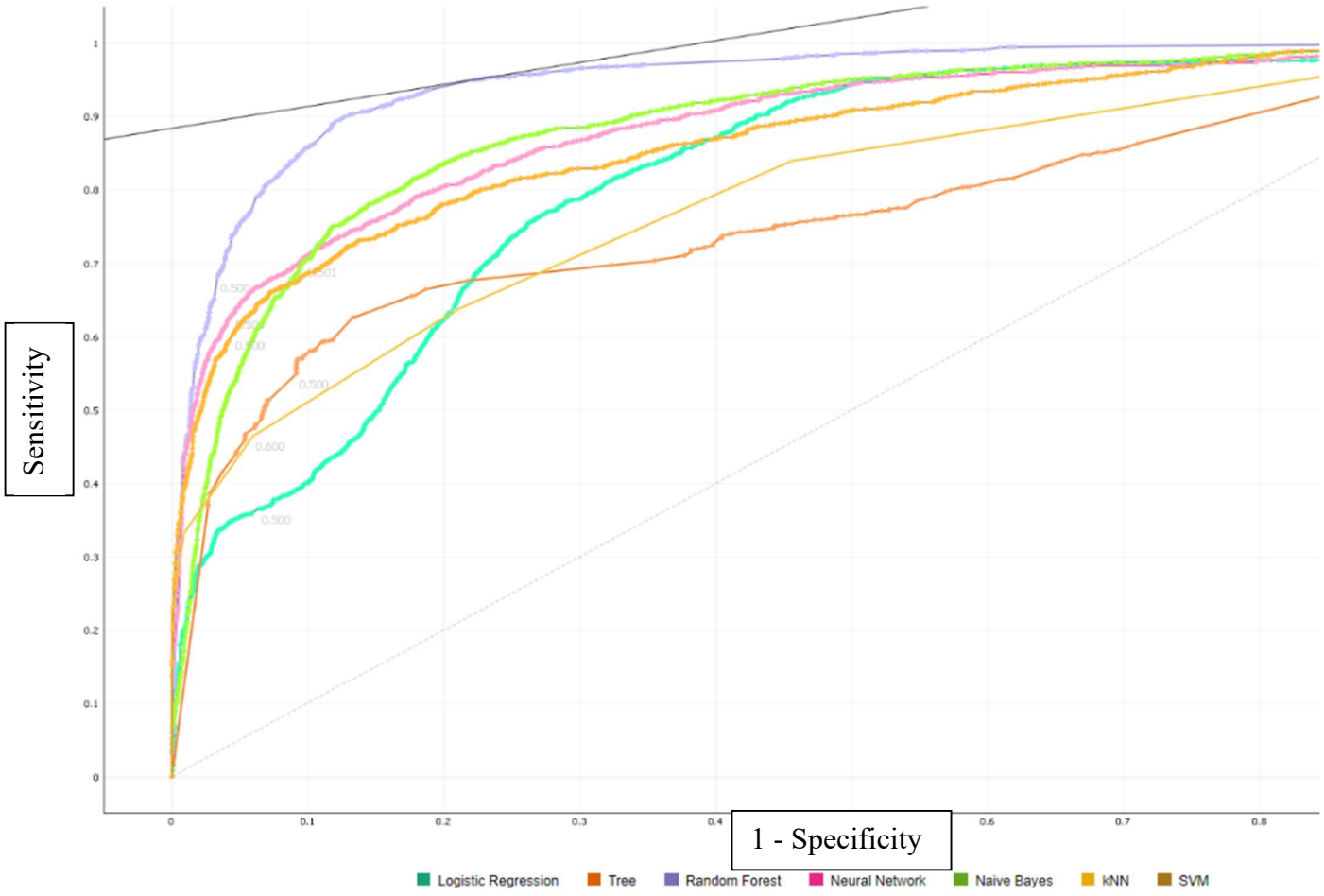


Figure 6-22: Lift curves for prediction models created in Orange

6.5.2. Analysis of Models Created in Orange

6.5.2.1. *Determining the Effect of Blank Cells on Accuracy*

Each workflow was created three times so the different datasets could be compared. The first workflow uses all the meeting minutes with unavailable information being left out or blank (called Dataset A). The second workflow uses all the meeting minutes where any missing information is substituted with a zero value in the relevant cell (called Dataset F). The third workflow was created from the selected meeting minutes of 12 projects (called Dataset S). The accuracy for the first two workflows is shown in Table 6-6. The test-and-score widget was set up to split the relevant test data (Dataset A, S or F) randomly into a training and testing dataset. The training and testing datasets are split 66% and 34% respectively. The model-training simulation was repeated 100 times, each time randomly splitting the dataset between training and testing.

The same dataset that was used to create the prediction model(s) was used to validate the prediction models. The prediction model is used to predict the outcome (in other words, the probability of late project completion). The model then compares the results to the actual values already contained in the dataset. The validation accuracies of the models created are shown in the last two columns of Table 6-6.. High accuracy results achieved for the prediction models indicate that the prediction models created were able to learn from the features that are present in the dataset and to use these features to predict the outcome of a project.

The test results show that the random forest learner created the most accurate model with an accuracy of 85.7%, followed by the neural network model with an accuracy of 85%. Using the results from the test-and-score widget, it was discovered that the accuracy of the created models is better when the cells are left blank (Dataset A) instead of being filled with an arbitrary constant (Dataset F). There was no change, however, in the type of learner that provided the best results.

Using the accuracy results for model validation, the accuracy of the models created were better for only three of the six models created when comparing the Datasets F and A.

The worst-performing learners in this study are the logistic regression learner and the kNN learner that respectively scored 66.1% and 63.1% accuracies overall. This, however, does not mean that they are not suitable or cannot be used. These accuracies may be high enough to provide some information.

Table 6-6: Accuracy scores for models created in Orange

Model type	Test and score		Validation	
	Accuracy Dataset F	Accuracy Dataset A	Accuracy Dataset F	Accuracy Dataset A
Random forest	85.7%	88.8%	97.9%	98.8%
Neural network	85%	85.3%	97.6%	96.9%
Tree	75.6%	78.6%	88%	84.9%
Naïve Bayes	70%	72.6%	74.4%	78.5%
Logistic regression	66.1%	66.5%	74.1%	70.1%
kNN	63.1%	73.8%	78.8%	84%

6.5.2.2. *Selected Meeting Minutes vs. All Meeting Minutes*

As indicated in Section 6.4.1 and Section 6.5.2.1, an additional dataset was created which contained minutes from 12 projects where the amount of information missing was less than 10% (Dataset S). When considering the use of the dataset with all the meeting minutes or the dataset with the selection of meeting minutes with the most information, the intention was to investigate two things:

- Whether or not the accuracy of the predictions change; and
- Whether the type of models that created the most accurate predictions would change.

By comparing the accuracy results from the two different datasets (A and S), it was discovered that all the accuracies increased when using the selected meeting minutes compared to when using all the minutes that are available. This is to be expected as a large amount of information is missing in Dataset A. This comparison also shows that, if more information is recorded in the meeting minutes, the models will inevitably become more accurate in their predictions. All the models returned accuracies above 80% on the more complete Dataset S. The accuracy of the best performing model in Section 6.5.2.1, the random forest model, increased by only 0.8%. This is almost negligible, considering that the accuracy is already 88.8%. It shows that this model can cope with the random nature of the meeting minutes and the lack of information in some meeting minutes. The changes in testing accuracy for this comparison is shown in Table 6-7. The type of models yielding the highest accuracies did not change.

Table 6-7: Accuracy of results for Datasets A and S using Orange

Model type	Accuracy for Dataset A	Accuracy for Dataset S
Random Forest	88.8%	89.6%
Tree	78.6%	83.1%
Neural Network	85.3%	87.9%
Logistic Regression	66.5%	80.4%
kNN	73.8%	82.8%
Naïve Bayes	72.6%	85.5%

6.5.3. Analysis of Models Created in RapidMiner

Figure 6-23 shows the typical output that can be obtained from each learner in RapidMiner. As an example, the output for the deep learning model is shown. This output also shows which features support the outcome for which the data analyst is looking. In this case, it was requested to show which features support an outcome where a project is late. According to the deep learning model, the features that support the prediction of late project completion are:

- Normalised site instructions (large impact);
- Progress planned;
- Rework mentioned;
- Design delays; and, to a lesser degree,
- Normalised client attendance.

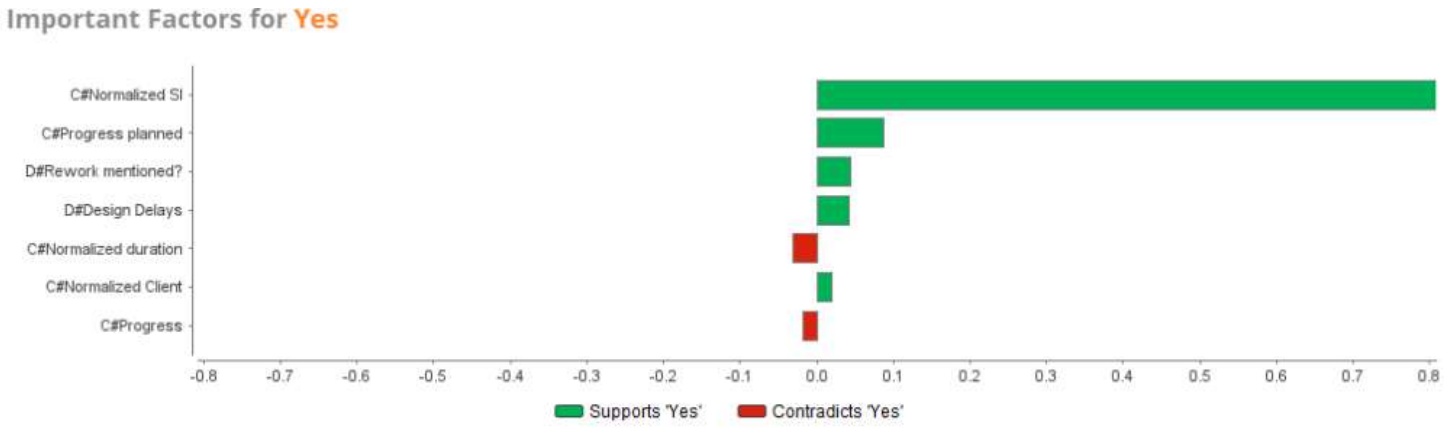


Figure 6-23: RapidMiner output of deep learning simulator

6.5.3.1. *Determining the Effect of Blank Cells on Accuracy*

The Auto Model feature in RapidMiner was used to create models from three different datasets (A, F and S). The accuracy of the models created from the two datasets (A and F) is shown in Table 6-8. The models' accuracy increased when the missing information was not added but rather left out. However, this was not so for the deep learning, decision tree and gradient boosted trees models. The models were created to predict the outcome of a project (in other words, whether the project will finish on time or late depending on the information contained in the dataset). The prediction models were tested in the same way by splitting the dataset into training and testing datasets, and repeating the process several times. From the analysis done, the best-performing model is created from the gradient boosted trees algorithm with an accuracy of 85%. This is followed by the random forest and deep learning algorithms with an accuracy of 81.7% and 79.2% respectively. The worst-performing algorithms are Naïve Bayes and fast large margin with 67.5% and 65.8% accuracy respectively.

Table 6-8: Accuracy for models created in RapidMiner

Model type	**Testing**	
	Accuracy of Dataset F	**Accuracy of Dataset A**
Gradient Boosted Trees	84.1%	85%
Random Forest	61%	81.7%
Deep Learning	74.7%	79.2%
Support Vector Machine	77.2%	73.3%
Decision Tree	63.7%	72.5%
Logistic Regression	70.5%	72.5%
Generalised Linear Model	66.4%	69.2%
Naïve Bayes	65.7%	67.5%
Fast Large Margin	78.6%	65.8%

6.5.3.2. *Selected Meeting Minutes vs. All Meeting Minutes*

All the models are more accurate when Dataset S is used compared to when Dataset A is used. The type of learner that created the most accurate model did change when comparing the two different datasets. Using Dataset S, the most accurate models created are the random forest and deep learning models, both with an accuracy of 88.3%. The most accurate model created for dataset A comes in third with an accuracy of 86.7% (gradient boosted trees). The

least accurate model is the SVM model with an accuracy of 75%. The accuracies for both Datasets A and S are shown in Table 6-9.

Table 6-9: Comparison of dataset S and A accuracies in RapidMiner

Model Type	Accuracy of Dataset A	Accuracy of Dataset S
Gradient Boosted Trees	85%	86.7%
Random Forest	81.7%	88.3%
Deep Learning	79.2%	88.3%
Support Vector Machine	73.3%	75%
Decision Tree	72.5%	83.3%
Logistic Regression	72.5%	78.3%
Generalised Linear Model	69.2%	78.3%
Naïve Bayes	67.5%	86.7%
Fast Large Margin	65.8%	81.7%

6.6. Refinement Process

As described in Section 3.3.1, the KDD process is cyclical in nature. Once the process has been completed, the entire process can be restarted. Slight changes can be made to the inputs, learner(s) and workflows to refine the model(s) being created. As part of the refinement process, features that enhanced the outcome of data mining were identified. Some of these features are discussed in Section 6.5.3. Listed below are the different iterations that were performed during the data mining process along with the changes made during the iterations.

1. Switch from applying all models to only some models. – This change was made as some of the learners are meant for linear analysis or text. Trying to apply all the learners to a single dataset in one workflow slowed the entire process down and, in some cases, stopped the data mining applications from working altogether.
2. Duplicate the dataset. – One set was left without change, while all blank spaces were filled with either “0”, or “nothing to report” in the other dataset. These sets are called Dataset A and F. This was done to determine how the learners handled not having all the features available and to find the best way in future to record missing information.
3. Split the numerical features from the text features in the dataset. – The new dataset is called Dataset T.

4. Refinement of the process and retesting to ensure that the obtained results are valid. – The workflows were updated to increase the testing and scoring parameters, and to do validation checks during the process to ensure that optimal models are being created. The stability of the prediction models created were verified, in other words the results between the iterations were consistent.

6.7. Lessons Learned from Applying the Data Mining Process

The different data mining applications provided models that are inherently different because each program uses different underlying algorithms. These differences extend to even the types of prediction models that provided the most accurate results. The best model offered by the Orange software for predicting the outcome of a project is the random forest model with an accuracy of 88.7%. The best model offered by the RapidMiner software is gradient boosted trees, with an accuracy of 85%. This did, however, change when using Dataset S. In this case, the most accurate model offered by the RapidMiner software is the one created from the random forest learner. The worst performing model offered by the Orange software was logistic regression. In RapidMiner, the worst-performing prediction model was the fast large margin. The type of model producing the best result is thus dependant on the software being used and the accuracy and completeness of the dataset being used. It is interesting that both data mining applications produced the most accurate models from Dataset S through the random forest learners. This might suggest that the random forest learner is the best algorithm to use for this type of data. However, it may change over time as more information is stored in the data warehouse, and more complete datasets are used to create the prediction models.

Creating more complex models did not always yield better results. As part of the refinement process, some changes were made to the input parameters of the learners. When tasked to create more complex models in RapidMiner, the computer-processing time went from a few minutes to several hours. The complex models created during this time did not yield any improvement over the models already created.

The accuracies and the most accurate models change depending on the information provided to the learners. It was found that Orange works better when the text and numerical features are split into different datasets. It allows for the application of learners best suited to each dataset. RapidMiner, on the other hand, can incorporate the text seamlessly, although different models are produced when the text is included or excluded. The speed of the

operation is also affected by the inclusion of the text. A normal execution of the application, without the text, would take a few minutes to complete. When text was included, the completion time could easily increase to hours.

6.8. Applying the Prediction Model(s)

The prediction models were applied to a newly-created dataset to provide an example where they are used to predict the outcome of a project. The new dataset, Dataset N, was created from meeting minutes that did not form part of the original Dataset A.

There are nine meeting minutes from two projects in the dataset. The first two sets of meeting minutes in the dataset are from Project 1 and the remaining meeting minutes in the dataset are from Project 2. The meeting minutes are from a specific point in the contract and their position in the dataset does not refer to their relative occurrence during the execution of the project. The first project was nearing completion while the second project had just started up. The predictions from these different models are shown in Figure 6-24, with the minutes from Project 1 and 2 highlighted.

Figure 6-24 shows the probability of a project completing behind schedule based on the information contained in the individual meeting's minutes. The probabilities are shown for seven of the prediction models created in Orange. The red-coloured bars are a visual indication of the magnitude of the prediction – in other words a prediction of 1.00 will have a red bar spanning the full width of the column indicating a 100% prediction that the project will complete behind schedule. The reverse is also true: where a prediction of 0.00 will not have a red bar and will have a prediction of 0% that the project will complete behind schedule.

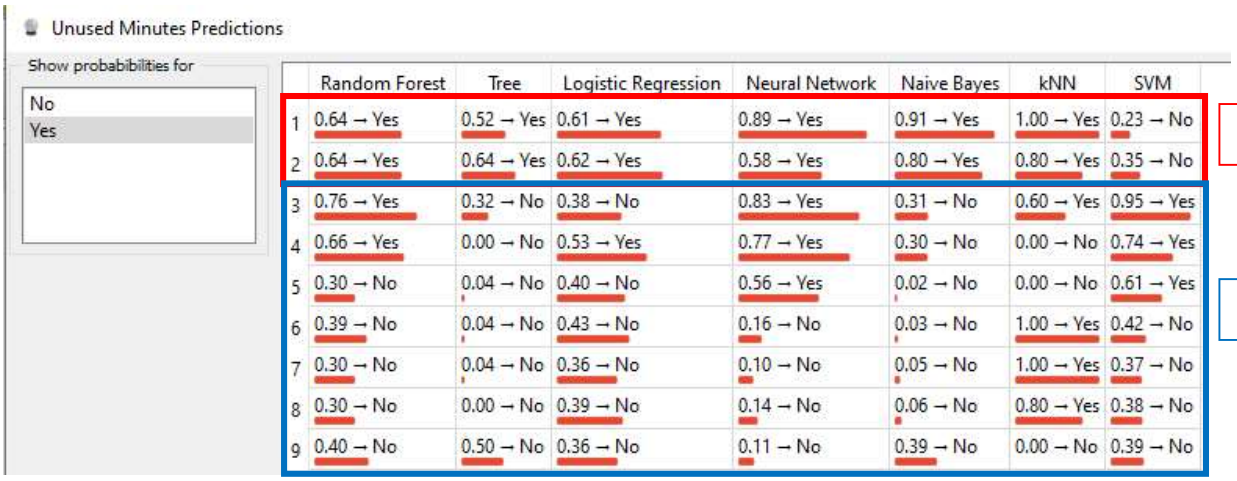


Figure 6-24: Predictions of project overrun from data mining models

Based on the information contained in the site meetings, six out of the seven models predict that the first project will complete behind schedule. However, the predictions for the second project are mixed, and the project has a good probability of running late if the project management team does not intervene. This coincides with the information from the meeting minutes where the first project is one month away from the completion date, but only about 92% complete and about 1% behind schedule.

The second project is around 8% and 19 days behind schedule by the 8th meeting. This project has approximately 12 months to go before the completion date. This means that it is possible to catch up. However, the trend indicates that the project loses about four days every month. Hence, some of the predictions from the models could be correct when showing about a 30% to 70% chance of the project completing behind schedule.

6.9. Conclusion and Discussion

The investigation of several projects' meeting minutes for the application of the KDD process followed the five steps of the KDD process. The conclusions that follow are in the same sequence, discussing each of the steps.

During the first step of the KDD process, the problem and expected results should be identified. The investigation uses project data (project meeting minutes) stored in a database to predict the outcome of a project. The result of the investigation is, that by storing large amounts of project data in a database or data warehouse, the information can be used in data mining applications to create prediction models to predict future outcomes.

During the second step of the KDD process, the data should be prepared for data mining. A new dataset was created from the meeting minutes of construction projects. The information contained in this dataset was manually entered into the relevant cells in a spread sheet by reading and transferring the information from each individual meeting's minutes. This process was not automated because of the variance in the format of the meeting minutes. There is no standardised format for the compilation of meeting minutes, how the information is recorded in the minutes or even what information is recorded in the minutes. This was a tedious and time-consuming step in the process. As part of the dataset creation, the data was transformed and arranged into the correct database structure for the next step. A proposed

standard for the recording of meeting minutes for future data mining is shown in Appendix 1 and is discussed in Chapter 7.

The third step in the KDD process is the mining of the dataset. Due to the novelty of using site meeting minutes as a dataset, two data mining applications were used to confirm the accuracy of the prediction models created and to serve as a secondary validation check on the models created. All the available data mining algorithms were used in both applications with eight algorithms in each data mining application. The original dataset (Dataset A) was copied to three other datasets, called Dataset F, S and T. Dataset F was changed by substituting an arbitrary zero value where information was missing. Dataset S is a shorter version of Dataset A including only the relevant entries that contained at least 90% of the information. Dataset T was created from Dataset A by removing all the numerical information, leaving only the text information in the dataset. The different data mining applications produced different results when using the same type of algorithms. This is because of the difference in the algorithms themselves and the differences in how they train, test and score the models.

The fourth step in the KDD process is to analyse the data that were obtained from the data mining process. The models that were created from Dataset F were not as accurate as the models created from Dataset A. The prediction models created from Dataset S were more accurate than the models created from Dataset A, proving that the accuracy of all the prediction models is increased if the meeting minutes include as much information as possible. It is thus important to ensure that all the information is recorded when the meeting minutes are recorded for future use. Every effort should be made by the project management team to ensure that the minutes are complete. Using the proposed standard template shown in Appendix 1 will assist in completing all the required information.

Text mining and sentiment analysis were also performed on Dataset T. No new knowledge could be obtained from mining the text. It was observed that there is a tendency for the sentiment of the meeting minutes to reflect the progress achieved on site.

A summary of the prediction models – named after the algorithms that were used – and their accuracies are shown in Table 6-10. It should be noted that a seasoned data analyst will know which models are better suited to the type of information contained in the minutes. The top two models created in both programs scored accuracies above 80%. It is interesting to note that three out of the four best-performing models are random forest models or variations thereof. This type of learner is well suited for the data mining of site meeting minutes.

Table 6-10: Summary of model accuracies

Prediction model	**Accuracy**
Random Forest (Orange)	88.8%
Neural Network (Orange)	85.3%
Gradient Boosted Trees (RapidMiner)	85%
Random Forest (RapidMiner)	81.7%
Deep Learning (RapidMiner)	79.2%
Tree (Orange)	78.6%
kNN (Orange)	73.8%
Support Vector Machine (RapidMiner)	73.3%
Naïve Bayes (Orange)	72.6%
Decision Tree (RapidMiner)	72.5%
Logistic Regression (RapidMiner)	72.5%
Generalized Linear Model (RapidMiner)	69.2%
Naïve Bayes (RapidMiner)	67.5%
Logistic Regression (Orange)	66.5%
Fast Large Margin (RapidMiner)	65.8%

The fourth and fifth steps are re-iterative steps where the data and new knowledge are analysed and subsequently refined where the discovered knowledge is validated. For this case study, no new knowledge could be discovered by data mining the datasets. However, the models that were created from the existing meeting minutes can predict the outcome of a project based on a limited number of meeting minutes – as little as one, if needed. However, a trend can appear if more than one meeting minutes are available for analysis when using the prediction models as seen in the example used for this study.

Overall, both data mining applications could produce models with high accuracies. The goal defined at the start of the KDD process will determine whether or not the models are suitable for use.

Applying the most accurate prediction model to Dataset N has shown that it is possible to predict (with some certainty) what the likely outcome of the project will be, based on the current information contained in the site's meeting minutes. Although a good project manager will have a "gut feeling" on where the project is headed based on their experience,

the data mining and analysis of the minutes are useful tools to prove and back up this "feeling". It can also be used to sway others to step up their performance and to address any shortcomings in the project execution.

In both applications, it is better to use the raw data without adding any missing information. The comparisons between the two datasets have shown that a more accurate model is created by leaving blank spaces instead of substituting arbitrary values in the empty spaces.

The correlations that were observed are all obvious correlations, such as the project duration correlating with the progress percentage. These correlations did, however, act as a verification that the information in the dataset is interpreted correctly by the data mining applications.

7. Proposal for Meeting Minutes

7.1. Introduction

Based on the results from the KDD process, site meeting minutes should be standardised as much as possible to facilitate the effective and efficient use of data mining. This chapter aims to provide a proposal and recommendation on the information that should be recorded in project meeting minutes to assist the KDD process.

7.2. Meeting Minutes Proposal

Numerical values or categorical values are better for using in the data mining process than text. However, meeting minutes need to have all these so that both humans and computers can use the information contained in the meeting minutes.

At the very least, the following items should be recorded in the meeting minutes. The **bold-face** items are the features that have been discovered to support the outcome of the project in Chapter 5. However, some of these items were not directly recorded in the minutes. They were calculated separately based on the information in the minutes for this study.

- Clear Target and Vision – The goal of the project should be stated together with when the project should be completed. The entire team should know why the project is needed and what the end goal is.
- **Progress achieved**, **progress planned** – State the progress percentage achieved to date and the progress percentage planned according to the latest revision of the program. The percentages should be based on earned value for the project. It is important that the percentages are calculated in the same way throughout the project.
- State whether the project is a**head/behind or on time** – Record the status of the project. This might differ from the progress planned. Record notes relating to the progress on site. It is vital to record the contract status consistently. If the float is being depleted or intermediate milestone dates are being missed, but the project is still predicted to complete on time, it should be stated as such.
- **Quality** – Record the quality of workmanship and include how many days the project has been delayed because of rework. The cost of rework should be

recorded as part of the contractor's quality management system. This includes the number of man-hours spent on the rework.

- **Field-engineering queries** – Record the number of FEQ's submitted to date. This should include whether any of the FEQ's submitted – within the timeframe of the meeting minutes – has an impact on the completion date of the project (for example, waiting for drawings).
- **Site instructions** – Record the total number of site instructions issued to the contractor to date.
- **Project completion date** – Record the anticipated and contractual project completion dates.
- Weather delays – Record the total number of weather delays that have been agreed upon.
- Project man-hours – Record the man-hours worked on the project to date. Where possible, the man-hours should be categorised into management, supervision, labour and subcontractors.
- Repeat items on the minutes – Record whether items have been carried over from the previous meeting.
- Record the total number of **compensation events** up to date.
- Record general information such as the **date** the meeting has taken place, **meeting number** and the attendees present at the meeting.

7.3. Create a database from the meeting minutes

This section describes the creation of a database and spread sheet that forms part of the proposed method of recording the meeting minutes. The use of these proposed methods will reduce the time needed to prepare the data for data mining.

Using software, in this case MS Access, an electronic form was created that contains all the attributes required from the meeting minutes. The form can be adjusted or customised to suit the individual projects by adding additional columns if more information needs to be recorded. Each time that minutes are typed into this electronic form, the information is automatically stored in a database, thus reducing the need to capture the information again when creating the data warehouse. A proposed form is shown in Appendix 1. Where additional information is attached to the meeting minutes, such as progress reports, the

important information from the attachment should be recorded in the meeting minutes. It should not be simply recorded as “see attached report”. That would ensure that the information is electronically available for data mining along with the remainder of the minutes in the data warehouse.

A separate spread sheet was set up in such a way that the database created from the electronic form can be exported to it. This spread sheet contains the formulas required for normalisation and the other calculations required for creating the dataset features. The spread sheet also contains the correctly formatted headers to assist with the process of loading the data into the data mining applications. This reduces the formatting time required by the data analyst before loading the data in the application. An example of this proposed spread sheet is attached in Appendix 5. The example does not contain all the headers for the dataset but provides an example of how the column headers should be formatted for Orange. The example also shows the formulas for the cells where calculations will be done based on the information from other cells. The assumption, when using this spread sheet, is that the information from the meeting minutes is stored chronologically.

MS Excel can import the data from this database directly for the data analyst to create the dataset. Since the headings are already in the correct format, all that the analyst has left to do is some checks on the contents to ensure that the information has been imported correctly before applying the chosen data mining model to the new dataset.

If Orange is used as the data mining application, this can all be done on-site by the project team at minimal cost. The process of data mining the meeting minutes and using the prediction algorithms to predict the outcome of a project can be used to guide the project-management team in their decisions taken on site.

8. Conclusion

Using the meeting minutes of a project for KDD is a novel concept. This study aimed to determine if the information currently being recorded in the site meeting minutes of a project is sufficient for use in data mining applications to predict the outcome of future projects.

The aim of the study was supported through four objectives. The outcome of each objective is discussed below.

8.1. Synthesis and Discussion of Project Success and Failure Factors

Objective 1 of this study is to identify factors that contribute to the success or failure of a project. In Chapter 2, an extensive literature review was performed to identify some of the main factors. Five research studies dealing with success factors in projects and five research studies dealing with failure factors were analysed and discussed. The main factors identified that are attributed to the success of a project are:

- Cost;
- Time management;
- Communication and consultation with stakeholders;
- Competent team members; and
- Goals and direction.

The main factors identified that are attributed to the failure of a project are:

- Leadership and governance;
- Planning; and
- Architecture and design.

The factors identified in the literature review can be grouped together under two main categories: leadership and planning. Table 8-1 indicates how these factors can be summarised in the categories. Some of the factors identified can be placed under both categories. The outcome of the literature review is that the leadership on a project has the biggest influence on its outcome. This is followed closely by planning, which is also related to the leadership on a project.

Table 8-1: Summary of the main two categories

Leadership	**Planning**
Time management	Time management
Communication and consultation with stakeholders	Planning
Architecture and design	Architecture and design
Goals and direction	Goals and direction
Leadership and governance	
Competent team members	

In Chapter 5, the two main success and failure categories are analysed to determine how the attributes could be recorded in the meeting minutes of a project. Some of the attributes can be used directly as features for data mining, while others need some refinement and calculations to create features that can be used in data mining. Based on the application of the prediction models that were created, the following success/failure attributes are important for data mining:

- Key performance indicators or goals of the project;
- Durations between site meetings (extracted from the dates of the meetings);
- Number of attendees at the meeting, both from the client's side and the contractor's side;
- Number of field engineering queries and site instructions;
- Quality comments;
- Repeating of items on minutes without them being addressed;
- Progress updates from the program, including being ahead or behind program;
- Number of compensation events (CE's) or claims from the contractor.
- Number of early warnings (NEC type contracts); and
- Project completion dates and the progress towards them.

8.2. Existing Data Mining and KDD in the Construction Industry

Objective 2 is to establish the status of data mining and KDD in the construction industry from literature.

Chapter 3 contains four research studies on the use of data mining in the construction industry. From these research studies, it is evident that data mining has been implemented in the construction industry. However, no evidence of data mining in the South African construction industry could be found. Some of the reasons identified why big data might not be utilised in construction projects include the organisation structure, incentive policies, top management support, resistance to change and exemplary projects.

8.3. Data Mining Meeting Minutes

Objective 3: Use the information obtained from Objective 1 to investigate if the attributes are being recorded in the meeting minutes of projects and how they can be turned into features.

The site meeting minutes are often recorded with different structures between different projects. From the meeting minutes that were used, it was found that much effort is needed to transform the information into a dataset that can be used for data mining. Using a standardised format for recording of meeting minutes can increase their efficiency in data mining. Chapter 7 proposes an electronic recording and storage process of minutes to automate the process of creating the features as much as possible.

Currently, the meeting minutes contain a large amount of information that can be stored as features in the data warehouse. Some attributes being recorded in the meeting minutes can be used as they are, while others need some modifications to become features. An example of the changes required is the normalisation of contract durations.

The two main categories that were identified as contributing towards the outcome of a project – leadership and planning – are being recorded in the meeting minutes. This was confirmed through the 10 attributes, identified in Chapter 5, that are recorded in most of the meeting minutes.

Objective 4: Use the KDD method to identify new and relevant information from the project meeting minutes to prove the value of the process.

The steps for using a KDD model are the following:

- Problem identification;
- Data preparation;
- Data mining;

- Data analysis and validation; and
- Refinement process.

The data mining of meeting minutes is presented in Chapter 6. Actual meeting minutes from various projects were used in two data mining applications to determine whether or not they can be used to predict the outcome of a project. The various data-preparation techniques and the large array of algorithms applied to the datasets are discussed. It is shown through analysis of the newly-created data mining models that the information currently being recorded in the project meeting minutes is sufficient to use to create prediction models.

Although no new knowledge was discovered during the KDD process, it proved that past meeting minutes could be used to predict the outcome of projects in progress. Once the prediction models are populated with more information from comprehensive meeting minutes, the prediction models will become more accurate. It is important that the information is recorded in a consistent manner to ensure optimal results during the KDD process.

It is thus shown that information recorded in construction project site meeting minutes is sufficient to be used in data mining applications to predict the outcome of future projects. It is shown that the KDD process can be used by project participants and the data mining approach can be used to benefit projects.

8.4. Recommendations for Further Research

This research focused on the meeting minutes of projects to data mine and create prediction models. The minutes of these projects normally form part of the official documentation available to all parties on the contract. There is a multitude of other documents, including emails and letters, that are generated on projects during the construction phase. It is recommended that all the electronic communication created for the project be stored in a data warehouse. The information in the data warehouse can then be transformed if needed for use in data mining applications, including text mining.

The meeting minutes used for this research often reported project progress as the percentage of work done on the total project. This is the most common way to report progress on a project. However, there are also other ways of reporting this. Alternative ways of reporting progress include:

- Comparing project expenditure versus planned expenditure – both as a total expenditure and as interim expenditure during the execution phase of the project; and
- Comparing progress being achieved on the critical path activities.

It is recommended that these alternative progress-report methods be investigated to determine the impact different reporting methods may have on the predictions.

Big data has much potential for industries willing to adapt to using it. One of the industries that are traditionally reluctant to change is the construction industry. It is recommended that a full-scale study be undertaken at a construction company, where all aspects of the recording and analysis of the data can be managed.

The information recorded in the site meeting minutes, or the lack thereof, could prove a valuable resource as this information may provide the client with the necessary insights to determine if the project management team in charge of a project is good or not. It is recommended that a study should be done on how this information can be used to determine the correlation between “good project management” and project completion (late or on time).

9. Bibliography

Alton, L., 2017. *The 7 Most Important Data Mining Techniques.* [Online]
Available at: https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques
[Accessed 23 June 2020].

Ananthanarayanan, K. & Cidrela Devi, A., 2017. Factors influencing cost over-run in Indian construction projects. *MATEC Web of Conferences.*

Arnx, A., 2019. *First neural network for beginners explained (with code).* [Online]
Available at: https://towardsdatascience.com/first-neural-network-for-beginners-explained-with-code-4cfd37e06eaf
[Accessed 30 July 2020].

Assaad, R., El-Adaway, I. H. & Abotaleb, I. S., 2020. Predicting Project Performance in the Construction Industry. *Journal of Construction Engineering and Management,* 146(5), pp. 1-22.

Association for Project Management, 2012. *APM Body of Knowledge.* Sixth ed. Buckinghamshire: Association for Project Management.

Association for Project Management, 2014. *Factors in Project Success,* Birmingham: BMG Research.

Beleiu, I., Crisan, E. & Nistor, R., 2015. Main Factors Influencing Project Success. *Interdisciplinary Management Research,* Volume XI, pp. 59-72.

Berg, R. G. v. d., 2020. *Logistic Regression - Simple Introduction.* [Online]
Available at: https://www.spss-tutorials.com/logistic-regression/
[Accessed 25 July 2020].

Botha, L. J., 2018. *Data Mining Construction Project Information to Aid Project Management.* Stellenbosch: University of Stellenbosch.

Burati, J. L., Farrington, J. J. & Ledbetter, W. B., 1992. Causes of Quality Deviations in Design and Construction. *Journal of Construction Engineering and Management,* 118(1), pp. 34-49.

Burke, R. & Barron, S., 2014. *Project Management Leadership.* Second ed. West Sussex: Wiley.

Calleam Consulting LTD, 2019. *101 Common Causes.* [Online]
Available at: http://calleam.com/WTPF/?page_id=2338
[Accessed 21 February 2019].

Cao, Y., Chau, K., Anson, m. & Zhang, J., 2002. An intelligent decision support system in construction management by data warehousing technique. *Lecture Notes in Computer Science,* Volume 2480, pp. 360-369.

Cardon, D., 2014. *Database vs. Data Warehouse: A comparative Review.* [Online]
Available at: https://www.healthcatalyst.com/database-vs-data-warehouse-a-comparative-review
[Accessed 28 March 2019].

Chan, D. W. N. & Kumaraswamy, M. M., 1997. A Comparative Study of Causes of Time Overrun on Hong Kong Construction Projects. *International Journal of Project Management,* 15(1), pp. 55-63.

Chauhan, N. S., 2019. *Decision tree Algorithm - Explained.* [Online]
Available at: https://towardsdatascience.com/decision-tree-algorithm-explained-83beb6e78ef4
[Accessed 26 July 2020].

Choudhary, A. K. et al., 2011. Knowledge discOvery And daTa minINg interGrated (KOATING) Moderators for collaborative projects. *International Journal of Production Research,* 49(23), pp. 7029-7057.

Chua, D. K. H., Kog, Y. C. & Loh, P. K., 1999. Critical success factors for different project objectives. *Journal of Construction Engineering and Management,* 125(3), pp. 142-150.

Dacombe, J., 2017. *An introduction to Artificial Neural Networks (with example).* [Online]
Available at: https://medium.com/@jamesdacombe/an-introduction-to-artificial-neural-networks-with-example-ad459bb6941b
[Accessed 26 July 2020].

Doloi, H., 2013. Cost Overruns and Failure in Project Management: Understanding the Roles of Key Stakeholders in Construction Projects. *Journal of Construction Engineering and Management,* 139(3), pp. 267-279.

Gaetsewe, R., Monyane, T. G. & Emuze, F., 2015. *Overruns Again in Public Projects: Perspective from Northern Cape, South Africa.* Brisbane, CEPPM.

Garbharran, H. & Govender, J., 2012. Critical success factors influencing project success in the construction industry. *Acta Structilia: Journal fo r the Physical and Developmental Sciences,* 19(2), pp. 90-108.

Hammad, A. & AbouRizk, S., 2014. Knowledge Discovery in Data: A Case Study. *Journal of Computer and Communications,* Volume 2, pp. 1-27.

Heravi, G. & Ilbeigi, M., 2012. Development of a comprehensive model for construction project success evaluation by contractors. *Engineering, Construction and Architectural Management,* 19(5), pp. 526-542.

Holgeid, K. & Thompson, M., 2013. A reflection on why large public projects fail. In: *The Governance of Large-Scale Projects - Linking Citizens and the State.* s.l.:The Hertie School of Governance and the Journal for Political Consulting and Policy Advice.

International Project Management Association, 2020. *Working Teams.* [Online]
Available at: https://www.apm.org.uk/body-of-knowledge/people/interpersonal-skills/leadership/
[Accessed 20 February 2020].

Koehrsen, W., 2017. *Random Forest Simplified.* [Online]
Available at: https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d
[Accessed 26 July 2020].

Kog, Y. C. & Loh, P. K., 2012. Critical Success Factors for Different Components of Construction Projects. *Journal of Construction Engineering and Management,* 138(4), pp. 520-528.

Lawrence, A. K., 2015. *Factors affecting project management services in Nigeria construction industry,* s.l.: s.n.

Marr, B., 2016. *Are these the 7 real reasons why tech projects fail?.* [Online]
Available at: https://www.forbes.com/sites/bernardmarr/2016/09/13/are-these-the-real-reasons-why-tech-projects-fail?
[Accessed 21 February 2019].

Moyce, C., 2014. Why do projects fail?. *Management Services,* 58(2), p. 40.

Nguyen, L. D., Ogunlana, S. O. & Lan, D. T. X., 2004. A study on project success factors in large construction projects in Vietnam. *Engineering, Construction and Architectural Management,* 11(6), pp. 404-413.

Nguyen, T. P. & Chileshe, N., 2014. Revisiting the construction project failure factors in Vietnam. *Built Environment Project and Asset Management,* 5(4), pp. 398-416.

Prabhakaran, S., 2018. *How Naive Bayes Algorithm Works? (with example and full code).* [Online]
Available at: https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/
[Accessed 27 July 2020].

Pratt, D., 2015. *Great Lessons in Project Management.* Tysons Corner: Management Concepts, Inc.

Project Management Institute, Inc., 2008. *A Guide to the Project Management Body of Knowledge (PMBOK Guide).* Fourth ed. Newtown Square: Project Management Institute, Inc..

Project Management Institute, 2016. *Construction Extension to the PMBOK Guide.* 3rd ed. Newtown Square: Project Management Institute, Inc..

Pupale, R., 2018. *Support Vector Machines (SVM) - An Overview.* [Online]
Available at: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989
[Accessed 26 July 2020].

RapidMiner, 2020. *RapidMiner Documentation.* [Online]
Available at: https://docs.rapidminer.com/
[Accessed 28 April 2020].

Ribeiro, P., Paiva, A., Varajao, J. & Dominuez, C., 2013. Success evaluation factors in construction project management - some evidence from medium and large Portuguese companies. *KSCE Journal of Civil Engineering,* 17(4), pp. 603-609.

Serrador, P., 2012. *The importance of the planning phase to project success.* Vancouver, Project Management Institute.

Sitwala, S. & Wium, J., 2013. *The Paradox of Time and Cost Overruns: A review of Literature in Developing Countries.* Cape Town, Association of Schools of Construction of Southern Africa.

Sohu, S. et al., 2018. Determining the Critical Success Factors for Highway Construction Projects in Pakistan. *Engineering, Technology & Applied Science Research,* 8(2), pp. 2685-2688.

Soibelman, L. & Kim, H., 2002. Data Preparation Process for Construction Knowledge Generation through Knowledge Discovery in Databases. *Journal of Computing in Civil Engineering,* 16(1), pp. 39-48.

Stecanella, B., 2017. *A Practical explanation of a Naive Bayes classifier.* [Online]
Available at: https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/
[Accessed 27 July 2020].

Stjepanovic, T. K., Tunc, A. O. & Janevska, J., 2020. *Academia.* [Online]
Available at:
https://www.academia.edu/37326378/Understanding_Leadership_in_International_Project_Management_Course.pdf
[Accessed 31 May 2020].

The Standish Group, 1994. *The Standish Group.* [Online]
Available at: https://www.standishgroup.com/sample_research_files/chaos_report_1994.pdf
[Accessed 21 July 2020].

Tshiki, M., 2015. Critical success factors for infrastructure construction projects in South Africa. *SAICE Civil Engineering Journal*, July, pp. 19-24.

University of Ljubljana, 2020. *Orange Documentation.* [Online]
Available at: https://orange.biolab.si/docs/
[Accessed 28 April 2020].

Valcheva, S., 2019. *Simple Linear Regression.* [Online]
Available at: http://www.intellspot.com/linear-regression-examples/
[Accessed 26 July 2020].

Windapo, A. & Akintona, R., 2015. *Establishing the relationship between construction project manager's skills and project performance.* College Station, Associated Schools of Construction.

Windapo, A., Odediran, S. & Akintona, R., 2015. *Establishing the relationship between construction project manager's skills & project performance.* s.l., Associated Schools of Construction.

Yu, T., Liang, X. & Wang, Y., 2020. Factors Affecting the Utilization of Big Data in Construction Projects. *Journal of Construction Engineering and Management,* 146(5), pp. 1-14.

Zahra, S. et al., 2014. Performing Inquisitive Study of PM Traits Desirable for Project Progress. *I.J. Modern Education and Computer Science,* Volume 2, pp. 41-47.

Zarsky, T. Z., 2003. "Mine Your Own Business!": Making The Case for the Implications of the Data Mining of Personal Information in the Forum of Public Opinion. *Yale Journal of Law and Technology,* 5(1), pp. 1-55.

Zhao, J., Pedrycz, W. & Senatore, S., 2013. *Data mining and knowledge discovery in industrial engineering.* [Online]
Available at:
http://link.galegroup.com/apps/doc/A378372012/AONE?u=27uos&sid=AONE&xid=0163c56e
[Accessed 7 April 2019].

Zulch, B., 2016. A proposed model for construction project management communication in the South African construction industry. *Acta Structilia,* 23(1), pp. 1-35.

Appendix 1: Meeting Minutes Template

Project Name	Meeting Number	Meeting Date
Project Vision/Target		
Client Attendees		
Contractor Attendees		
Health, Safety and Environmental		
Security/IR		
Quality		
Contractual/Commercial		
Progress Planned (%)	Progress Achieved (%)	Days Ahead (+) or Behind (-)

Progress Notes		
Project Completion Date	Planned Completion Date	Project Man-hours
Number of FEQ's	Number of SI's	Design Delays
Early Warnings: Client	Early Warnings: Contractor	Early Warnings" Resolved
Compensation Events		
Weather Delays		

Appendix 2: Dataset Example

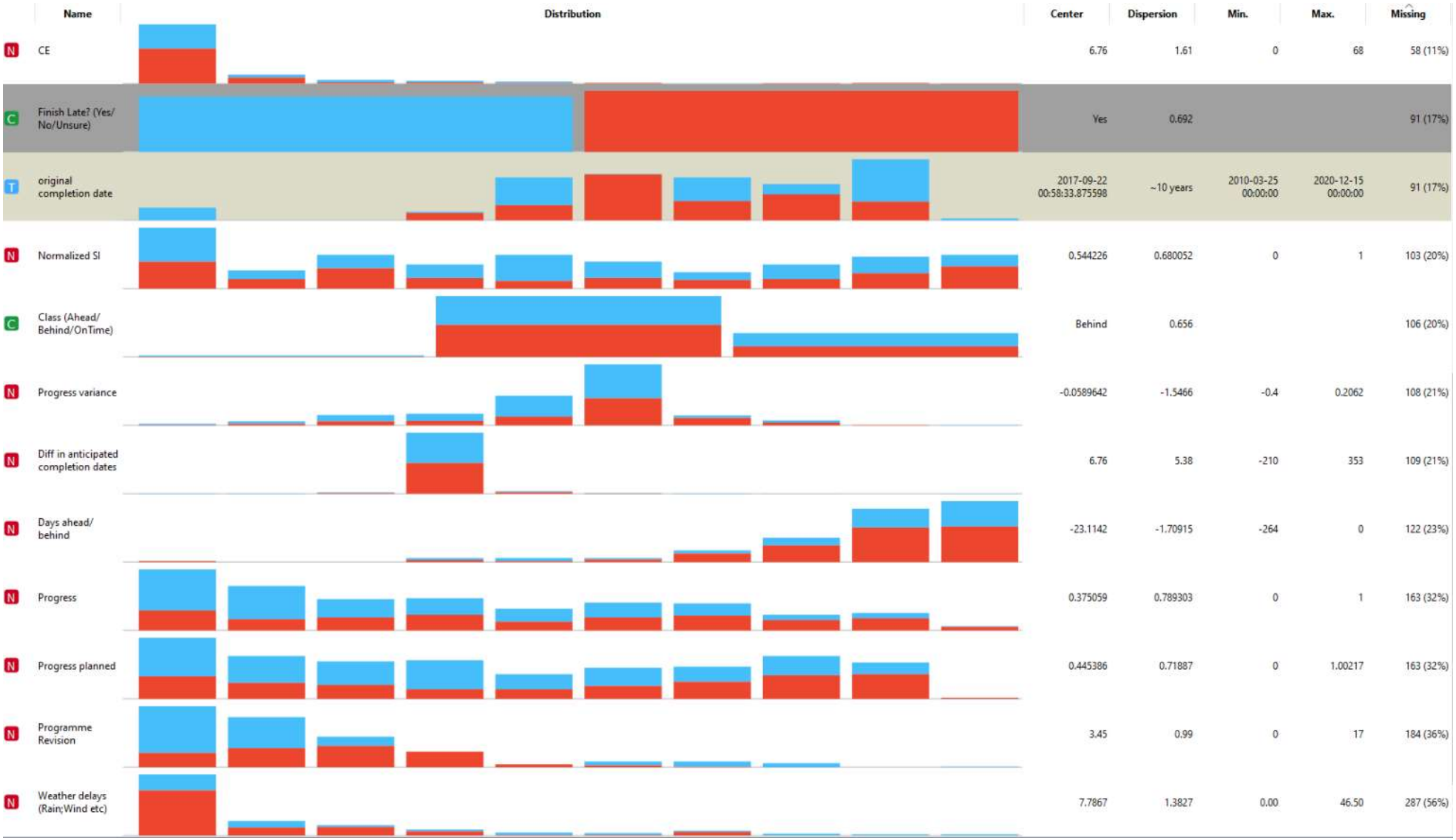
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	C#Meeting Number	iT#Date	C#Contractors	C#Client	mS#Health, Safety, Enviro & Community	mS#Security/IR	mS#Quality	mS#Contractual/Commercial	C#Progress	C#Progress planned	iT#Expected completion date	mS#Progress Notes	C#No of TQ's	C#No of SI's	C#duration of contract elapsed	mT#original completion date	i#EW Contractor	i#EW Client	C#CE	C#Days ahead/ behind	i#Project Manhours	C#Progress variance	C#Days between Meetings	cD#Class (Ahead/Behind/On Time)	C#Weather delays (Rain; Wind etc)	C#Diff in anticipated completion dates	D#Rework mentioned?
2	0	20-Jun-14	0	0	to	Nothing t	Nothing	to	0%	0%	20-Oct-14	to	0	0	0	20/October/2014	0	0	0	0	0	0.00%	0	OnTime	0	0	No
3	1	4-Jul-14	3	5	r's	kers	enquire	ual	7%	10%	41932	ti Socks	0		14	20/October/2014			0	-5		-3.00%	0	Behind		0	No
4	2	17-Jul-14	2	4	Safety	kers	enquire	ual	15%	21%	41932	issue	1		27	20/October/2014			0	-8		-6.00%	13	Behind		0	No
5	3	31-Jul-14	2	4	from	plants	leave a	suggest	24%	32%	41932	issue	1		41	20/October/2014			0	-12		-8.00%	14	Behind		0	No
6	4	12-Aug-14	3	2	from	Steelwor	leave a	and AC	27%	46%	41932	issue	1		53	20/October/2014			0	-15		-19.00%	12	Behind		0	Yes
7	5	26-Aug-14	2	3	team	Steelwor	submit	and AC	36%	61%	41932	issue	1		67	20/October/2014			0	-15		######	14	Behind		0	Yes
8	6	9-Sep-14	2	3	team	Steelwor	submit	and AC	56%	71%	41932	issue	7		81	20/October/2014			0	-15		-15.00%	14	Behind		0	No
9	7	23-Sep-14	2	3	Health &	kers	submit	and AC	67%	75%	41932	issue	8		95	20/October/2014			0	-10		-8.00%	14	Behind		0	No
10	8	21-Oct-14	2	2	Health &	Nothing	updated	and AC	75%	84%	41960	issue	8		123	17/November/2014			0	-10		-9.00%	28	Behind		28	Yes
11	9	4-Nov-14	2	3	Health &	Nothing	updated	and AC	84%	94%	41960	issue	10		137	17/November/2014			0	-10		-10.00%	14	Behind		0	Yes
12	10	18-Nov-14	2	3	to report	to report	to report	to report	89%	100%	41960	to report	10		151	17/November/2014			0	-10		-11.00%	14	Behind		0	Yes
13	11	2-Dec-14	2	4	Health &	Nothing	updated	and AC	94%	100%	41985	issue	10		165	22/November/2014			0	-10		-6.00%	14	Behind		25	Yes
14	12	13-Jan-15	2	3	Health &	Nothing	updated	and AC	82%	83%	42061	issue	10		207	26/February/2015			0	-4		-1.00%	42	Behind		76	No
15	13	27-Jan-15	1	3	Health &	to	updated	and AC	91%	92%	42039	issue	12		221	26/February/2015			0	-6		-1.00%	14	Behind		-22	Yes
16	14	24-Feb-15	2	4	Health &	Nothing t	b.Ran	C.ITEM and AC	100%	100%	07/March/2015	issue	13		249	26/February/2015			0	-6		0.00%	28	Behind		31	Yes
17	0	19-Jan-16	4	3	accepta	Security		ACTUAL	0%	0%	18/January/2017	to report	0	0	0	18/January/2017	0	0	0	0	0	0.00%	0	OnTime	0	0	No
18	1	23-Feb-16	5	3	Accom	Security	CATION	to report	6%	8%	18/January/2017	Program		9	35	18/January/2017			0	0		-2.56%	35	OnTime	0	0	No
19	2	5-Apr-16	3	5	CATION	second	e quality	to report	18%	21%	18/January/2017	on		19	77	18/January/2017			0	-10		-2.74%	42	Behind	4	0	No
20	3	25-Apr-16	4	4	NMENT	Security	RD OF	to report	18%	27%	18/January/2017	32 days		26	97	18/January/2017			0	-32		-8.77%	20	Behind	6	0	Yes
21	4	24-May-16	5	4	keeping	Contract	discussi	to report	26%	35%	18-Jan-17	to report		30	126	18/January/2017			2	-30		-8.22%	29	Behind	7	0	No
22	5	21-Jun-16	4	4	keeping	Contract	discussi	to report	34%	42%	13-Mar-17	to report		37	154	25/January/2017			2	-30		-8.22%	28	Behind	9	54	No
23	6	27-Jul-16	4	4	keeping	Contract	done	to report	38%	51%	5-Apr-17	to report		45	190	25/January/2017			4	-48		-12.90%	36	Behind	11	23	No
24	7	23-Aug-16	4	4	keeping	Tempor	discussi	to report	44%	58%	13-Apr-17	to report		50	217	25/January/2017			5	-54		-14.52%	27	Behind	11	8	No
25	8	27-Sep-16	3	4	keeping	Tempor	Enginee	to report	48%	68%	11-May-17	to report		64	252	25/January/2017			6	-74		-19.89%	35	Behind	11	28	No
26	9	25-Oct-16	3	4	keeping	Tempor	Enginee	to report	55%	75%	15-May-17	to report		75	280	25/January/2017			7	-76		######	28	Behind	13	4	Yes
27	10	29-Nov-16	3	4	Enginee	Enginee	discussi	to report	62%	85%	26-May-17	to report		79	315	25/January/2017			7	-83		-22.31%	35	Behind	17	11	Yes
28	11	18-Jan-17	3	4	keeping	reinforci	e Works	to report	60%	79%	29-May-17	to report		88	365	23/April/2017			7	-88		-19.13%	50	Behind	20	3	No
29	12	21-Feb-17	3	4	keeping	Warning	discussi	to report	83%	87%	9-May-17	to report		99	399	24/April/2017			7	-15		-3.25%	34	Behind	28	-20	No
30	13	22-Mar-17	2	4	House ke	The AoT	The defe	Nothing t	89%	93%	18-May-17	Nothing to report		101	428	24/April/2017			8	-16		-3.47%	29	Behind	36	9	Yes
31	14	25-Apr-17	4	4	ouse kee	The AoT	The finish	Nothing t	92%	100%	6-Jun-17	Nothing to report		111	462	24/April/2017			8	-26		-8.65%	34	Behind	36	19	No
32	15	30-May-17	3	4	House ke	During th	finishing	Nothing t	94%	100%	3-Jul-17	Nothing to report		114	497	24/April/2017			9	-50		-6.23%	35	Behind	36	27	No
33	16	27-Jun-17	4	4	Theft on	House ke	Kruger	Nothing t	96%	100%	10-Jul-17	Nothing to report		114	525	24/April/2017			9	-55		-4.03%	28	Behind	36	7	No
34	17	26-Jul-17	2	5	House ke	Theft on	Edwin Kr	Nothing t	98%	100%	14-Aug-17	Nothing to report		115	554	24/April/2017			9	-80		-1.83%	29	Behind	36	35	No
35	18	22-Aug-17	3	4	Discussi	No discus	Concrete	Nothing	100%	100%	42972	to report		115	581	24/April/2017			9	-87.86		0.00%	27	Behind	36	11	No
36	0	16-May-17	5	2	Medical	Nothing t	9.1 The	Contract	0%	0%	27-Oct-17	or to	0	0	0	27-Oct-17	0	0	0	0	0	0.00%	0	OnTime	0	0	No
37	1	20-Jun-17	3	2	Existing	Nothing t	SS	Contract	0%	0%	31-Oct-17	to report			35	31-Oct-17				0		0	35	OnTime	0	0	No
38	2	1-Aug-17	2	3	use of	Nothing t	concret	date	0%	0%	31-Oct-17	to report			77	31-Oct-17				0		0	42	OnTime	0	0	No
39	3	22-Aug-17	2	2	of safety	Nothing t	concret	date	0%	0%	31-Oct-17	submitte			98	31-Oct-17				0		0	21	OnTime	1	0	No
40	4	7-Sep-17	3	2	use of	Nothing t	concret	date	0%	0%	31-Oct-17	late			114	31-Oct-17				-4		0	16	Behind	1	0	No
41	5	21-Sep-17	3	2	There	Nothing t	Mahiken	date	0%	0%	31-Oct-17	late			128	31-Oct-17				-6		0	14	Behind	6	0	No
42	6	12-Oct-17	2	2	There	Nothing t	Afrisam	date	0%	0%	31-Oct-17	wind			149	31-Oct-17				0		0	21	OnTime	6	0	No
43	7	26-Oct-17	3	3	There	Nothing t	quality	date	0%	0%	31-Oct-17	actual			163	31-Oct-17				0		0	14	OnTime	6.5	0	No
44	0	29-Aug-16	0	0	contract	Nothing t	Nothing	a.Steran	0%	0%		EKLY	0	0	0		0	0	0	0	0	0.00%	0		0	0	No
45	1	20-Sep-16	3	5	the first	to report	submitte	mention				submitte			22								22				No
46	2	11-Oct-16	4	6	ion	to report	y	ct sign-				submit			43								21				No
47	3	18-Oct-16	3	5	indicate	to report	followed	RE				complet			50								7				No
48	4	25-Oct-16	4	6	informed	to report	followed	RE				complet			57								7				No
49	5	1-Nov-16	4	4	& SS-	to report	followed	RE				t			64								7				No

Appendix 3: Visualisation of Dataset Features

The features-statistics widget in the Orange data mining application provides the user with a way to inspect the dataset prior to starting the data mining process. It is meant to provide the user with some interesting features of the dataset. The target feature was used for the feature statistics shown in Figure A-1. In this case, the target feature is the categorical feature "Project Late" with the two categories – Yes/No. The two differently coloured bars in the figure show the proportions in each bin for Yes (red) and No (blue). There are eight columns that show the statistics of the dataset. Each of the columns are described below, from left to right:

1. The feature type is determined by the user before the dataset is created. The feature type can be **C**ategorical, **N**umeric, **T**ime or **S**tring. The letters in the coloured blocks indicate the feature type. In this case, there are only two feature types in the selection – Categorical and Numerical.
2. The name of the feature.
3. Histograms of the feature values. If the feature is numeric, the application discretises the values into bins to create the histograms. If the values are categorical, each category is assigned its own bar in the histogram. The histograms provide the user with an indication of the distributions of each feature within the dataset.
4. The central tendency of the feature values as calculated by the application. If the feature is numerical, the mean or average of the feature is shown. If the feature is categorical, the mode or the value that appears the most often is shown.
5. The dispersion of the features is shown. For numerical features, the coefficient of variation or relative standard deviation is calculated. For categorical features, the dispersion is the entropy or average level of uncertainty of the feature.
6. The minimum value is only calculated for numerical and ordinal categorical features.
7. The maximum value is only calculated for numerical and ordinal categorical features.
8. The missing number of values for the specific feature is shown both in number and a percentage of the total.

RapidMiner has similar functions for statistical analysis of the dataset. Figure A-2 and Figure A-3 are screenshots from the RapidMiner application showing the feature statistics for the dataset. The statistics shown are similar to Orange with the distributions of the feature, minimum, maximum, average and standard deviation being shown.



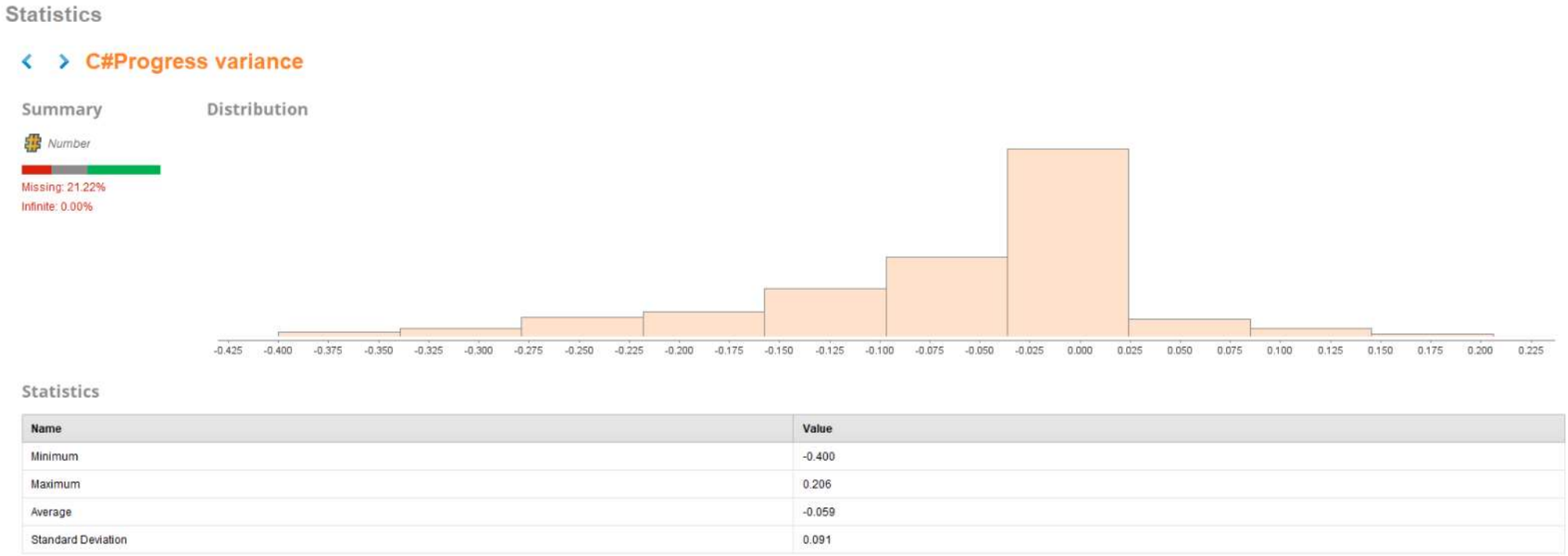
	Name	Distribution	Center	Dispersion	Min.	Max.	Missing
N	CE		6.76	1.61	0	68	58 (11%)
C	Finish Late? (Yes/ No/Unsure)		Yes	0.692			91 (17%)
T	original completion date		2017-09-22 00:58:33.875598	~10 years	2010-03-25 00:00:00	2020-12-15 00:00:00	91 (17%)
N	Normalized SI		0.544226	0.680052	0	1	103 (20%)
C	Class (Ahead/ Behind/OnTime)		Behind	0.656			106 (20%)
N	Progress variance		-0.0589642	-1.5466	-0.4	0.2062	108 (21%)
N	Diff in anticipated completion dates		6.76	5.38	-210	353	109 (21%)
N	Days ahead/ behind		-23.1142	-1.70915	-264	0	122 (23%)
N	Progress		0.375059	0.789303	0	1	163 (32%)
N	Progress planned		0.445386	0.71887	0	1.00217	163 (32%)
N	Programme Revision		3.45	0.99	0	17	184 (36%)
N	Weather delays (Rain;Wind etc)		7.7867	1.3827	0.00	46.50	287 (56%)

Figure A-1: Dataset features visualisation from Orange application

Data

D#Finish Lat... Category	C#CE Number	C#Days ahea... Number	C#Days betw... Number	C#Diff in anti... Number	C#Normalize... Number	C#Normalize... Number	C#Normalize... Number	C#Normalize... Number	C#Normalize... Number	C#normalize... Number	C#Programm... Number	C#Progress Number	C#Progress p... Number	C#Progress v... Number
Yes	0	0	0	0	0	0	0	0	0	0	?	0	0	0
Yes	0	-5	0	0	1.667	1.500	0.056	0.071	?	0	?	0.070	0.100	-0.030
Yes	0	-8	13	0	1.333	1	0.108	0.143	?	0.077	?	0.150	0.210	-0.060
Yes	0	-12	14	0	1.333	1	0.165	0.214	?	0.077	?	0.240	0.320	-0.080
Yes	0	-15	12	0	0.667	1.500	0.213	0.286	?	0.077	?	0.270	0.460	-0.190
Yes	0	-15	14	0	1	1	0.269	0.357	?	0.077	?	0.360	0.610	-0.250
Yes	0	-15	14	0	1	1	0.325	0.429	?	0.538	?	0.560	0.710	-0.150
Yes	0	-10	14	0	1	1	0.382	0.500	?	0.615	?	0.670	0.750	-0.080
Yes	0	-10	28	28	0.667	1	0.494	0.571	?	0.615	?	0.750	0.840	-0.090
Yes	0	-10	14	0	1	1	0.550	0.643	?	0.769	?	0.840	0.940	-0.100
Yes	0	-10	14	0	1	1	0.606	0.714	?	0.769	?	0.890	1	-0.110
Yes	0	-10	14	25	1.333	1	0.663	0.786	?	0.769	?	0.940	1	-0.060
Yes	0	-4	42	76	1	1	0.831	0.857	?	0.769	?	0.820	0.830	-0.010
Yes	0	-6	14	-22	1	0.500	0.888	0.929	?	0.923	?	0.910	0.920	-0.010
Yes	0	-6	28	31	1.333	1	1	1	?	1	?	1	1	0
Yes	0	0	0	0	0.750	1.333	0	0	0	0	?	0	0	0
Yes	0	0	35	0	0.750	1.667	0.060	0.056	0.078	?	?	0.058	0.083	-0.026

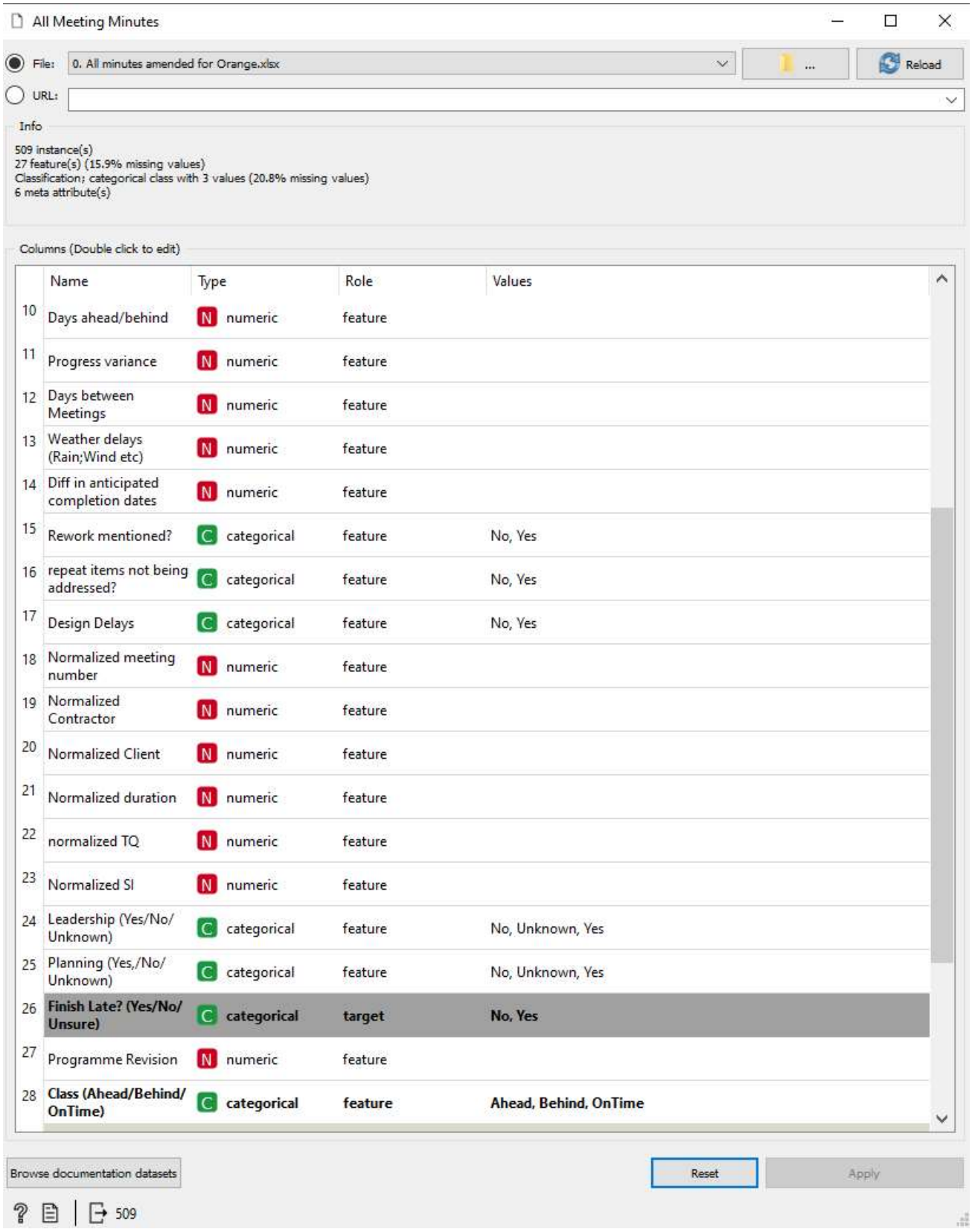
Figure A-2: Dataset features visualisation from RapidMiner



Name	Value
Minimum	-0.400
Maximum	0.206
Average	-0.059
Standard Deviation	0.091

Figure A-3: Statistics for Progress Variance feature from RapidMiner

Appendix 4: Data Features Role and Type Allocation



	Name	Type	Role	Values
10	Days ahead/behind	numeric	feature	
11	Progress variance	numeric	feature	
12	Days between Meetings	numeric	feature	
13	Weather delays (Rain;Wind etc)	numeric	feature	
14	Diff in anticipated completion dates	numeric	feature	
15	Rework mentioned?	categorical	feature	No, Yes
16	repeat items not being addressed?	categorical	feature	No, Yes
17	Design Delays	categorical	feature	No, Yes
18	Normalized meeting number	numeric	feature	
19	Normalized Contractor	numeric	feature	
20	Normalized Client	numeric	feature	
21	Normalized duration	numeric	feature	
22	normalized TQ	numeric	feature	
23	Normalized SI	numeric	feature	
24	Leadership (Yes/No/ Unknown)	categorical	feature	No, Unknown, Yes
25	Planning (Yes,/No/ Unknown)	categorical	feature	No, Unknown, Yes
26	**Finish Late? (Yes/No/ Unsure)**	**categorical**	**target**	**No, Yes**
27	Programme Revision	numeric	feature	
28	**Class (Ahead/Behind/ OnTime)**	**categorical**	**feature**	**Ahead, Behind, OnTime**

			C#Meeting Number
			iT#Date
			C#Contractors
			mS#Quality
			C#Progress
			C#Progress Planned
			iT#Expected Completion Date
			mS#Progress Notes
			C#No of FEQ's
			C#No of SI's
=B4-start date of project	=B3-start date of project	=B2-start date of project	C#Duration of Contract Elapsed
			mT#Original Completion Date
=I4-J4	=I3-J3	=I2-J2	C#Progress Variance
=B4-B3	=B3-B2	=B2-previous meeting date	C#Days between Meetings
=P4-P3	=P3-P2	=P2-project completion date	C#Diff in Anticipated Completion Dates
Yes/No	Yes/No	Yes/No	D#Rework Mentioned?
Yes/No	Yes/No	Yes/No	D#Repeat Items Not Being Addressed?
Yes/No	Yes/No	Yes/No	D#Design Delays
=A4/MAX(A$3:A$101)	=A3/MAX(A$3:A$101)	=A2/MAX(A$3:A$101)	C#Normalised Meeting Number
=C4/ROUND(AVERAGE(C$3:C$100);0)	=C3/ROUND(AVERAGE(C$3:C$100);0)	=C2/ROUND(AVERAGE(C$3:C$100);0)	C#Normalised Contractor
=D4/ROUND(AVERAGE(D$3:D$100);0)	=D3/ROUND(AVERAGE(D$3:D$100);0)	=D2/ROUND(AVERAGE(D$3:D$100);0)	C#Normalised Client
=IF(MAX(O$3:O$100)=0;0;O4/MAX(O$2:O$100)	=IF(MAX(O$3:O$100)=0;0;O3/MAX(O$2:O$100))	=IF(MAX(O$3:O$100)=0;0;O2/MAX(O$2:O$100))	C#Normalized Duration
=IF(MAX(M$2:M$99)=0;0;M4/MAX(M$2:M$100	=IF(MAX(M$2:M$99)=0;0;M3/MAX(M$2:M$100))	=IF(MAX(M$2:M$99)=0;0;M2/MAX(M$2:M$100))	C#normalised FEQ
=IF(MAX(N$3:N$100)=0;0;N4/MAX(N$2:N$100)	=IF(MAX(N$3:N$100)=0;0;N3/MAX(N$2:N$100))	=IF(MAX(N$3:N$100)=0;0;N2/MAX(N$2:N$100))	C#Normalised SI
Yes/No	Yes/No	Yes/No	D#Finished Late?